THE COMPATIBILITY OF THE STRUCTURE-AND-DYNAMICS ARGUMENT AND PHENOMENAL FUNCTIONALISM ABOUT SPACE

BY

LUKE ROELOFS

Abstract: Chalmers argues against physicalism using the premise that no truth about consciousness is deducible *a priori* from purely structural truths, and later defines what it is for a truth to be structural, which turns out to include spatiotemporal truths. But Chalmers then defines spatiotemporal terms by reference to their role in causing spatiotemporal experiences. Stoljar and Ebbers argue that these definitions allow for the trivial falsification of Chalmers premise about structure and consciousness. I show that this result can be avoided by tweaking the relevant premise, and that this tweak is not *ad hoc*.

Daniel Stoljar and Melissa Ebbers have recently argued that two key parts of the metaphysical framework developed by David Chalmers are incompatible. On the one hand, this framework is committed to a principled conceptual separation between consciousness, understood as 'non-structural', and physics, understood as 'structural'. We can formulate this as follows:

Non-Entailment (NE): No truth about consciousness follows *a priori* from any set of purely structural truths.

This principle, together with the claim that physics can teach us only structural truths, serves both as an argument against physicalism (Chalmers,

Pacific Philosophical Quarterly •• (2017) ••–•• DOI: 10.1111/papq.12184

2002, 2010) and also as a principled explanation of what underlies the intuitive plausibility of other arguments against physicalism, such as the conceivability argument and the knowledge argument (see esp. Alter, 2016). Philosophical zombies are conceivable because their full physical description is purely structural and so does not entail any consciousness; Mary does not know what it's like to see red because her complete physical knowledge is purely structural and so does not entail knowledge of consciousness.

But to understand Non-Entailment we need some understanding of the notion of 'structure'. While there are several ways to understand this (see Alter and Nagasawa, 2012; McClelland, 2013; Pereboom, 2014; Alter, 2016; cf. Stoljar, 2014, 2015, pp. 7–13), Chalmers' way is to define a structural truth as one that is *a priori* equivalent to one formulable in a certain restricted set of terms, specifically 'spatiotemporal expressions, nomic expressions, and mathematical and logical expressions' (2010, p. 120 n17). Thus NE says we cannot reach conclusions about consciousness if we start from premises formulated in only these terms.

The difficulty concerns the meaning of spatiotemporal expressions (though for simplicity I will henceforth focus just on spatial terms). We might understand these as ascribing primitive, undefinable, spatial properties, or as ascribing properties definable in terms of some more basic properties (e.g. causal structure). But Chalmers (2012, pp. 325–336) argues that the best way to understand them is what he calls 'phenomenal functionalism'. This says that 'to a first approximation, our concepts of [spatial properties] pick out those properties that normally bring about certain spatial experiences and judgments' (p. 327). Phenomenal functionalism is preferable to its rivals because, on the one hand, it far from obvious that space can be entirely defined in more basic structural terms like causal or mathematical structure (cf. Chalmers, 2012, p. 332; Stoljar, 2015, pp. 16–17; Alter, 2016, pp. 810–814), and because, on the other hand, treating spatial properties as *sui generis* has the sceptical implication that for all we know the world might in fact be devoid of spatial properties, with some other properties causing us to have spatial experiences (cf. Chalmers, 2012, pp. 333–334).

However, Stoljar claims that a contradiction arises from simultaneously endorsing NE, defining 'structural' to include 'spatial', and defining 'spatial' by reference to experience.

> [The definition of 'structural' implies] that physical truths are equivalent to various truths formulated in a certain vocabulary including spatiotemporal vocabulary; and phenomenal functionalism says that truths formulated in spatiotemporal vocabulary are a priori equivalent to (or a priori entail) truths about consciousness. Putting this together… physical truths are a priori equivalent to, or a priori entail, truths about consciousness (Stoljar, 2015, p. 15).

How is this supposed to work? Here is a first pass (simpler than what Stoljar ends up saying) at the problematic entailment:

1. There is an object *x*, with spatial properties.
2. To have spatial properties is to have properties which are the normal causes of spatial experience.
3. *Therefore*, *x* has properties which are the normal causes of spatial experience. (from 1 and 2)
4. If something has the properties which are the normal cause of some event, then that event has or will happen at least once.
5. *Therefore*, spatial experience has or will happen at least once. (from 3 and 4)

We seem to have derived a truth about consciousness (5) from a purely structural truth (1) together with some definitions (2 and 4). This would falsify Non-Entailment. However, claim 4 is false: something need not cause what it is the normal cause of. For instance, suppose certain mushrooms are the normal cause of a certain distinctive human experience; nevertheless those mushrooms might exist, with all their chemical properties, in a world where humans never evolved.

Because properties might be the normal cause of some effect, and yet exist in a world where that effect never comes about, we must specify which possible worlds are being talked about. The point of NE is that from structural facts about some world, no phenomenal facts about that world follow. Thus, a proper formulation of the above inference, which makes clear that 4a is false, would be:

1a There is an object *x* (in world *w*) with spatial properties.
2a To have spatial properties (in any world) is to have properties which are (in the actual world @) the normal causes of spatial experience.
3a *Therefore*, *x* has (in *w*) properties which are (in @) the normal causes of spatial experience. (from 1a and 2a)
4a If something has (in *w*) the properties which are (in @) the normal cause of some event, then that event has happened or will happen at least once (in *w*).
5a *Therefore*, spatial experience has happened or will happen at least once (in *w*). (from 3a and 4a)

Stoljar recognises this difficulty, but suggests that it can be repaired by replacing the false general claim 4a with a contingent structural claim, as follows:

1a There is an object *x* (in world *w*) with spatial properties.
2a To have spatial properties (in any world) is to have properties which are (in the actual world @) the normal causes of spatial experience.
3a *Therefore*, *x* has (in *w*) properties which are (in @) the normal causes of spatial experience. (from 1a and 2a)

4b The properties of object *x* cause (in *w*) what they are (in @) the normal causes of.[1]

5a *Therefore*, spatial experience has or will happen at least once (in *w*). (from 3a and 4b)

This entailment still threatens NE as long as 4b is itself a purely structural truth. And it appears that it is – it contains causal terms ('causes', 'cause of'), logical terms ('The properties of object *x*'), and the term 'normally'. This term is the hardest to analyse, but it seems to mean either something strictly statistical ('As are the normal causes of Bs' = 'the majority of Bs are caused by As'), or else something about the dispositions of As ('As are the normal causes of Bs' = 'As have a robust tendency to cause Bs under a range of circumstances'), or some combination of the two.[2] Since statistical terms plausibly count as 'logical/mathematical', and dispositional terms plausibly count as 'causal/nomic', 4b seems to qualify as structural whatever we think 'normal cause of' means.[3] Thus from purely structural truths (1a and 4b), together with a definition (2a), we can derive a truth about consciousness (5a). Hence NE is false.

Melissa Ebbers (Ebbers, n.d.) has argued that this incompatibility ramifies further, impugning the conceivability argument directly, even though this argument is usually framed not in terms of 'structural' expressions but in terms of 'physical' properties. Ebbers' argument is that Chalmers' favoured way of defining physical properties (2012, pp. 319–321), just like his favoured way of defining spatiotemporal properties, involves an ineliminable reference to experiences.[4] In short, terms for properties like 'being an electron' are *a priori* equivalent to 'whatever property has caused a certain pattern of observations *via* a certain mathematically-specified structure of interaction with other properties'. Since those 'observations' involve experiences, Ebbers claims that philosophical zombie worlds turn out not to be conceivable after all (n.d., pp. 7–11). Her reasoning is that since the physical truths about some world are defined ultimately by reference to experiences, they can only be true if there are some experiences to give them meaning. In a zombie world there are no experiences, hence physical terms are meaningless, and so it cannot be true that the zombies are physically identical to us. Hence from the fact that the zombies are physically identical to us, it follows *a priori* that there must be experiences in that world, and thus that it is not a zombie world.[5] This threatens Chalmers' argument because he seems to accept:

Conceivability from A Priority (CAP): P& ~ Q is conceivable if and only if P does not entail Q *a priori.*

Given CAP, and Ebbers' argument that physical truths *a priori* entail certain experiences, it follows that zombie worlds (where physical truths obtain without any experiences) are not conceivable. But that conceivability claim is the first premise of the conceivability argument against physicalism.

Although I think Stoljar's and Ebbers' arguments succeed with NE and CAP defined as above, I believe that properly-qualified forms of NE and CAP undercut them, and moreover that these qualified forms are well-motivated, simply serving to more explicitly capture the spirit of the original NE and CAP.[6] To begin to see this, observe the special role played in Stoljar's argument by the actual world: we are able to draw conclusions about experience because we are able to leverage the fact that it is by our actual-world experiences that we fix reference to properties. The spirit of NE, it seems to me, is not threatened by the possibility of this kind of manoeuvre: it is meant to rule out things like a functional reduction of consciousness, which would permit functional facts about some world to entail phenomenal facts about that same world. But finding an explicit formulation of this idea is tricky. Here is my proposal:

Qualified Non-Entailment (QNE): No truth about consciousness in a world *w*, considered as counterfactual, is *a priori* entailed by any set of purely structural truths concerning only the world *w*.[7]

This principle would rule out 4b ('The properties of object *x* cause (in *w*) what they are (in @) the normal causes of'), because that premise relies on comparing the behaviour of *x*'s properties in the two worlds *w* and @. Thus the fact that 5a follows from 1a and 4b is not a counterexample to QNE. Put colloquially, the point is simply this: phenomenal functionalism says that 'space' is whatever causes certain experiences *here*. To know that they cause those experiences *somewhere*, we would need the premise that they have the same effects *there* as they do *here*, and QNE specifically disregards any employment of such 'cross-world comparisons'. QNE says simply that no amount of structural information about things over *there* can entail consciousness *there*.

Two phrases in QNE deserve unpacking. First, it is restricted to worlds 'considered as counterfactual' (as opposed to worlds 'considered as actual'). This distinction, drawn from Davies and Humberstone, 1980, is central to Chalmers' modal framework, and so appropriate to employ here. To consider a world as actual, 'one thinks of a possibility as representing a way the actual world might turn out to be', while to consider a world as counterfactual, 'one acknowledges that the actual world is fixed, and thinks of a possibility as a way the world might have been but is not' (Chalmers, 2004, p. 159).[8]

Second, QNE mentions 'purely structural truths concerning only the world *w*.' This cannot simply mean 'purely structural truths which are true only 'at' the world *w*', since pretty much any truth true at one world will also be true at various others. Rather, it means 'purely structural truths which are true solely in virtue of how things are in world *w*.' This rules out 'cross-world comparisons', since their truth-conditions contain (at least) one requirement

that something be true at one world, and (at least) one requirement that something be true at a different world. That is (to use 4b as an example) what matters is not which worlds 4b gets an assignment of truth relative to, but whether its truth depends on both *w* and @ being certain ways.

It might be thought that Stoljar's objection could be resurrected by simply running the argument with an exclusive focus on the actual world: if something has spatial properties in the actual world, there must be spatial experiences in the actual world. I think an argument like this can work; indeed, I think an even simpler argument also works, namely that if anything in any world has spatial properties, then there must be spatial experiences in the actual world – if there were none then our spatial terminology would be meaningless (assuming phenomenal functionalism is correct). But this does not provide a counter-example to QNE, for the argument relies on treating the actual world as actual, i.e. allowing it to determine the meanings of certain terms, whereas QNE specifies that worlds are to be considered as counterfactual.

Alternatively, we might try to resurrect Stoljar's argument by talking about a disposition to cause spatial experiences, as follows:

1a There is an object *x* (in world *w*) with spatial properties.
2a To have spatial properties (in any world) is to have properties which are (in the actual world @) the normal causes of spatial experience.
3a *Therefore*, *x* has (in *w*) properties which are (in @) the normal causes of spatial experience. (from 1a and 2a)
4c If something has (in *w*) the properties which are (in @) the normal cause of some event, then those properties confer (in any world) a disposition to cause that event.
5c *Therefore, x* has (in *w*) a certain disposition D which is (in @) manifested, and whose manifestation involves causing spatial experience. (from 3a and 4c)
6c Object *x* manifests disposition D (in *w*).
7c *Therefore*, spatial experience has or will happen at least once (in *w*). (from 5c and 6c)

Here 4c is a plausible construal of the meaning of 'normal cause', and 6c is offered as a purely structural truth concerning only world *w*. If 7c follows, we would thus seem to have a counterexample to QNE. However, 6c is not in fact purely structural, since its use of the name 'D' is not *a priori* equivalent to anything that can be said in exclusively causal, nomic, mathematical, etc. terms. D is picked out specifically by reference to its role in causing spatial experience in the actual world, so any description of it that can be substituted *a priori* will have to *mention* spatial experience in the actual world. To put it another way, the argument does not go through *a priori* unless we build into the meaning of the name 'D', not just that it is

a disposition bestowed by possession of spatial properties, but also that it is manifested in the actual world. The latter fact stops this from being a counter-example to QNE.

The response to Ebbers' argument is similar: qualify CAP so as to disregard *a priori* entailments that rely on the fact that a certain world is actual:

Qualified Conceivability from A Priority (QCAP): P& ~ Q is conceivable if and only if P does not entail Q *a priori* when all worlds are considered as counterfactual.[9]

This rescues the conceivability of zombie worlds, because even though physical truths *a priori* entail the existence of consciousness in the actual world (to give meaning to physical terms), this entailment cannot be drawn when we are considering all worlds as counterfactual – the zombie world can be conceived of as a counterfactual possibility, which we describe using physical terms whose meaning is fixed in the actual world, not in the world which is being conceived of. This world contains properties that, in our own world, cause certain kinds of experiences, but which does not do so in the zombie world: the attribution of physical properties is true because we are speaking a language whose meanings are fixed by the actual world, not by the worlds we are using it to describe.

We might still worry that replacing NE with QNE, and CAP with QCAP, seems like an ad hoc fix. But we can see that it is not when we see that, in general, *any* sort of claim about the failure of one kind of truth to a *priori* entail another kind of truth, and any claim about a dissociation between two things being conceivable, will need to be qualified in the manner of QNE and QCAP. Without such a qualification, a counterexample can be constructed with sufficient ingenuity.

For instance, suppose we wish to say the following: simply from truths about the pattern of instantiation of static shape properties, we cannot *a priori* deduce truths about biology, or about happiness, or about ethics. This is surely true, if any claim of this sort is. Correlatively, it is surely true that any pattern of shapes could conceivably exist without life, happiness, or rights. But now suppose that in the actual world, a certain shape is most often instantiated by happy tortoises, who have a moral right not to be needlessly harmed. Suppose that we name that shape 'the T-shape', making it *a priori* that the T-shape is, in the actual world, most often instantiated by happy tortoises, who have a moral right not to be needlessly harmed. Now suppose we are considering a possible world *w*, about which we know only the following:

Suspicious Premise (SP): The T-shape is, in *w*, instantiated by what it is, in the actual world, most often instantiated by.

This appears to be simply a claim about the instantiation of static shape properties – that a certain pattern of similarity holds between the instantiations of *this* particular shape in *w* and in @. But it follows *a priori* that there is, in *w*, at least one tortoise, and at least one happy thing, and at least one thing with rights. Thus we have falsified our original claim that shapes are not enough to entail biology, happiness, or ethics. To falsify the claim that shapes without life, happiness, or rights are conceivable, we similarly appeal to CAP and to the *a priori* entailment from 'the T-shape is instantiated by what it is most often instantiated by' to 'there are happy tortoises with moral rights'. Both of these moves are clearly too easy: this sort of *a priori* entailment is not what was ever at issue. And a way to rule it out is to rule out cross-world premises like SP, and to base conceivability claims only on entailments which treat no world as actual. In most cases, when we are interested in which claims *a priori* entail which other claims, we are really interested in which claims *about some world* can *a priori* entail which other claims *about that world*. When this point is recognised, the inconsistencies in the Chalmersian framework identified by Stoljar and Ebbers disappear.

School of Philosophy
Australian National University

NOTES

[1] Note that saying something 'causes what it is the normal cause of' is different from saying it 'causes what it normally causes.' If A normally causes B, that means it is rare to find A without B; if A is the normal cause of B, that means it is rare to find B without A. Spatial properties are the normal causes of spatial experience, but do not normally cause spatial experience: most instances of them are unperceived and so do not cause any experiences.

[2] Here is a reason for thinking that robust dispositions matter: even if it turned out that most blue-ish experiences in human history had been caused by capricious gremlins who telepathically induced blue-ish experiences whenever a human being was around something reflecting light in the 'red' spectrum, wouldn't we still say that the really blue things are not those which reflect 'red' light, but rather those which a human would see as blue in the absence of telepathic meddling? Here is a reason for thinking that statistics matter: if it turned out that certain objects reliably and directly produce blue-ish experiences by releasing a gas that caused synaesthesia, the question of whether they or light-reflecting objects are 'blue', or simply 'blue-seeming' will turn on which ones have in fact caused any, all, or most, of the blue-ish experiences in human history.

[3] A slightly different possibility is that 'the normal cause of' means 'the most frequent cause of under specified 'normal' conditions': this would force us to ask what makes certain conditions 'normal', with the same three options of statistical regularity, dispositional robustness, or a combination of the two, re-appearing.

[4] More precisely, physical terms are mostly theoretical terms (with the possible exception of a few, like 'mass', see Chalmers, 2012, pp. 322–324), and hence will be 'Ramsified' into descriptions of their relations to each other, and to 'Observational terms', which include among other things terms reporting observer's experiences.

[5] One response available to Chalmers is that even if a zombie world is inconceivable, worlds with some zombie inhabitants are not – there are enough experiences to provide meaning to the

physical terms used to characterise the zombies. This would still seem just as potent an objection to physicalism, which is committed to global entailment, but it also compromises the clarity of Chalmers' general argument and goes against his explicit claims, so I will assume that this fall-back position is unacceptable to him.

[6] Chalmers' own response is to say that the entailments identified by Stoljar and Ebbers are unproblematic so long as they do not give rise to metaphysical entailments. I do not intend my position to conflict with this, but to appeal simply to metaphysical entailment would undermine the dialectical force of the arguments where metaphysical entailment (or its absence) is the conclusion, and a *priori* entailment (or its absence) is the premise. So it is necessary to explicitly qualify those doctrines which are usually expressed simply in terms of a priori entailments, which is what I seek to do.

[7] This could be put more formally as follows: for any world *w*, and any truth about consciousness Q, and any set of purely structural truths P, then as long as we consider *w* as counterfactual, it is never a *priori* that if P obtains in *w*, then Q obtains in *w*.

[8] To see the difference, consider the possibility that all the samples we have identified as water are and always have been composed of a substance chemically nothing like H2O, and scientists have somehow misanalysed them. Considering this world as actual, we would have to say that water is not, after all, H2O: the two have turned out to be distinct. But considering the world as counterfactual, we should instead describe it as a case where human beings have never encountered water (that is, H2O), but some other substance.

[9] This could be put more formally as follows: P& ~ Q is conceivable if, for any world *w* considered as counterfactual, it is not a *priori* that if P obtains in *w*, then Q obtains in *w*.

REFERENCES

Alter, T. (2016). "The Structure and Dynamics Argument against Materialism," *Noûs* 50(4), pp. 794–815.

Alter, T. and Nagasawa, Y. (2012). "What Is Russellian Monism?" *Journal of Consciousness Studies* 19(9–10), pp. 67–95.

Chalmers, D. (2002). "Consciousness and its Place in Nature," in D. Chalmers (ed.) *Philosophy of Mind: Contemporary and Classical Readings*. New York: Oxford University Press, pp. 247–272.

Chalmers, D. (2004). "Epistemic Two-Dimensional Semantics," *Philosophical Studies* 118(1–2), pp. 153–226.

Chalmers, D. (2010). *The Character of Consciousness*. New York: Oxford University Press.

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

Davies, M. and Humberstone, L. (1980). "Two Notions of Necessity," *Philosophical Studies* 38(1), pp. 1–31.

Ebbers, M. (n.d.). 'A Priori Entailment and the Reference-Fixing Problem.'

McClelland, T. (2013). "The Neo-Russellian Ignorance Hypothesis: A Hybrid Account of Phenomenal Consciousness," *Journal of Consciousness Studies* 20(3–4), pp. 125–151.

Pereboom, D. (2014). "Russellian Monism and Absolutely Intrinsic Properties," in U. Kriegel (ed.) *Current Controversies in Philosophy of Mind*. London: Routledge, pp. 40–69.

Stoljar, D. (2014). "Four Kinds of Russellian Monism," in U. Kriegel (ed.) *Current Controversies in Philosophy of Mind*. London: Routledge, pp. 1–39.

Stoljar, D. (2015). "Russellian Monism or Nagelian Monism?" in T. Alter and Y. Nagasawa (eds) *Consciousness in the Physical World: Perspectives on Russellian Monism*. New York: Oxford University Press, pp. 324–345.