Is Consciousness Intrinsic? A Problem for the Integrated Information Theory

Hedda Hassel Mørch, University of Oslo

Published in *Journal of Consciousness Studies*, Volume 26, Numbers 1-2, 2019, pp. 133-162(30), special issue on the Integrated Information Theory (ed. Garrett Mindt)

Penultimate draft – for citation please use published version.

Abstract: The Integrated Information Theory of consciousness (IIT) claims that consciousness is identical to maximal integrated information, or maximal Φ. One objection to IIT is based on what may be called the intrinsicality problem: consciousness is an intrinsic property, but maximal Φ is an extrinsic property; therefore, they cannot be identical. In this paper, I show that this problem is not unique to IIT, but rather derives from a trilemma that confronts almost any theory of consciousness. Given most theories of consciousness, the following three claims are inconsistent. INTRINSICALITY: Consciousness is intrinsic. NON-OVERLAP: Conscious systems do not overlap with other conscious systems (*a la* Unger's problem of the many). REDUCTIONISM: Consciousness is constituted by more fundamental properties (as per standard versions of physicalism and Russellian monism). In view of this, I will consider whether rejecting INTRINSICALITY is necessarily less plausible than rejecting NON-OVERLAP or REDUCTIONISM. I will also consider whether IIT is necessarily committed to rejecting INTRINSICALITY or whether it could also accept solutions that reject NON-OVERLAP or REDUCTIONISM instead. I will suggest that the best option for IIT may be a solution that rejects REDUCTIONISM rather than INTRINSICALITY or NON-OVERLAP.

1 Introduction

The Integrated Information Theory (IIT) of consciousness claims that consciousness is identical to maximal integrated information, or maximal Φ (Tononi et al. 2016; Tononi et al. 2014). IIT has gained by now a significant following, based on a combination of empirical and philosophical evidence. But IIT is also subject to a number of philosophical objections. One of these objections is that consciousness intuitively seems like an intrinsic property, but maximal Φ is an extrinsic property. Therefore, if our intuitions are correct, consciousness cannot be identical to maximal Φ (Chalmers 2016: 200, fn 8). An intrinsic property is a property that does not constitutively depend (though it may well causally depend) on properties of other things, or its external surroundings. An extrinsic property is as a property that *does* constitutively depend on properties of other things, or its external surroundings. Intuitively, one cannot change whether someone is conscious merely by changing the properties of other things (assuming these changes do not causally affect the conscious area of their brain), but IIT seems to imply that one can—most obviously in virtue of its *Exclusion* postulate. IIT's *Exclusion* postulate claims that consciousness requires *maximal* Φ, as opposed to merely some non-zero value of Φ, and whether a system has maximal Φ depends on whether it is subsumed by any larger system with higher Φ, which would be an external circumstance. Call this *the intrinsicality problem.*

Prima facie, the intrinsicality problem is one of the most serious philosophical objections that face IIT. Other objections typically target the philosophical evidence for IIT (more specifically, its argument from axioms to postulates—this will be discussed in more detail below) (Cerullo 2015; Bayne 2018), rather than IIT itself, and could thereby only justify reducing one's credence in IIT to what is merited by the empirical evidence alone, which would arguably still be significant. Other philosophical objections to IIT claim, in effect, that it fails to solve the hard problem of consciousness (Mindt 2017), or alternatively, that is fails to avoid standard objections to physicalism or other metaphysical views about the hard problem compatible with IIT. However, this objection can also be made against any other neuroscientific theories of consciousness. The intrinsicality problem, in contrast, threatens to show that an identity between maximal Φ and consciousness is impossible, and would remain impossible even if the hard problem (or objections to identity-based solutions to it) is set aside.[1]

In this paper, I will show that the intrinsicality problem is actually not unique to IIT, but rather derives from a trilemma that confronts almost any theory of consciousness. According to the trilemma, given any plausible theory of consciousness—including IIT's most important rivals such as the Global Workspace Theory—the following three claims are inconsistent:

> INTRINSICALITY: Consciousness is intrinsic.
>
> NON-OVERLAP: Conscious systems do not overlap with other conscious systems (*a la* the problem of the many (Unger 1980))
>
> REDUCTIONISM: Consciousness is constituted by more fundamental properties (as per standard versions of physicalism and Russellian monism).

A version of this trilemma has previously been raised by Trenton Merricks (1998), but not applied to neuroscientific theories of consciousness. Peter Unger (2004), Eric Schwitzgebel (2015) and Jonathan Simon (2018) have raised important aspects of the trilemma for neuroscientific theories, but not the full version. I will argue that the full trilemma holds for all plausible neuroscientific theories.

This means that rejecting IIT in favor of some other theory will not help preserve INTRINSICALITY unless one is also willing to give up either NON-OVERLAP or REDUCTIONISM. In defense of IIT, one might therefore argue that rejecting INTRINSICALITY is no less (or perhaps even more) plausible than rejecting NON-OVERLAP or REDUCTIONISM. One might also consider whether IIT really is committed to rejecting INTRINSICALITY rather than NON-OVERLAP or REDUCTIONISM, or insofar as

[1] I am not claiming that the intrinsicality problem is the only problem of this kind (and thus *the* most, as opposed to one of the most, serious philosophical objections to IIT). Another problem of the same kind, which is also more well-known, is Scott Aaronson's (2014) objection that structures (such a expanders and grids) that intuitively have nothing to do with consciousness can have maximal and very high Φ. This paper does not address this objection.

it is, whether this commitment can be abandoned without any harmful consequences for the rest of the theory.

I will examine the viability of all these responses and conclude that, in general, rejecting INTRINSICALITY is not clearly less plausible than rejecting NON-OVERLAP. Rejecting NON-OVERLAP is also possible for IIT, by abandoning the *Exclusion* postulate. However, this will not fully ensure compatibility between IIT and INTRINSICALITY because, as I will discuss, there is a sense in which Φ itself is also extrinsic. Rejecting REDUCTIONISM, on the other hand, would ensure full compatibility with INTRINSICALITY, but many would consider REDUCTIONISM more plausible than INTRINSICALITY, so this might seem like a bad option. However, I will argue that, for IIT in particular, the cost of rejecting REDUCTIONISM is lower than one might expect. The best option for IIT all things considered might therefore be to keep INTRINSICALITY and NON-OVERLAP and reject REDUCTIONISM.

2 The Integrated Information Theory and the Intrinsicality Problem

IIT's central claim is that a physical system is conscious if and only if it is a maximum of integrated information, or Φ for short. Integrated information, as IIT defines it, can be roughly described as a measure of the extent to which a system causally constrains its own past and future state, together with the extent to which these constraints depend on the causal interconnectivity between the system's parts. A system is a *maximum* of integrated information if it has more integrated information (i.e. is more self-constraining in virtue of its internal interconnectivity) than any overlapping system, i.e., any smaller system that is part of it and any larger system that it is a part *of*.[2] The claim that consciousness requires maximal Φ, as opposed to just some non-zero value of Φ, is known as the *Exclusion* postulate.[3] IIT also claims that consciousness is *identical* to maximal Φ.[4]

According to the intrinsicality problem, it is highly counterintuitive that consciousness is identical to maximal Φ, because consciousness intuitively seems like an intrinsic property but maximal Φ is an extrinsic property. This objection has been briefly raised by David Chalmers (2016: 200, fn 8), but has otherwise not been discussed in any detail.

As noted, an intrinsic property can be defined as a property of a system or entity that does not constitutively depend on properties of other things, or on what is going on in its external

[2] Note that being part of a system in IIT's sense requires functionally contributing to it (i.e., that removing the part affects the functioning of the containing system) rather than merely spatially overlapping with it.
[3] For a more elaborate introduction to IIT, see Tononi and Koch (2015). For the full technical details, see Tononi et al. (2014)
[4] IIT specifically claims that a conscious experience is identical to a "maximally irreducible conceptual structure" (MICS). Maximal irreducibility is equivalent to having maximal Φ and the conceptual structure refers to the specific causal structure that has maximal Φ, so this seems roughly equivalent to asserting an identity between maximal Φ and consciousness.

surroundings.[5] An extrinsic property can be defined as a non-intrinsic property, or a property of a system or entity that *does* constitutively depend on properties of other things, or what is going on in its surroundings. An example of an intrinsic property is height because someone's height cannot be changed merely by changing their surroundings. An example of an extrinsic property is "being the tallest person in the room" because this can be changed merely by changing the person's surroundings, i.e., by adding or removing other tall people to or from the room.

Consciousness[6] intuitively seems like an intrinsic property: whether someone is conscious does not seem to depend on external circumstances in anything like the way "being the tallest person in the room" does. This is of course not to say that consciousness cannot *causally* depend on external circumstances. For example, if oxygen is removed from our surroundings, or a poisonous gas is added, we will quickly lose consciousness. But this is because lack of oxygen, or presence of a poisonous gas, *causes* damage to the brain which is in turn directly responsible for loss of consciousness. If (by some miraculous intervention) oxygen could be removed from, or poisonous gas be added to, our environment without causally affecting the brain, it would not lead to loss of consciousness. Intuitively, no purely extrinsic changes, i.e., changes to the environment of a conscious system that do not cause changes within the conscious system itself, can change the consciousness of a system.

Maximal Φ, in contrast, is an extrinsic property. Whether a system has *maximal* Φ depends on whether the system is part of any larger system that has higher Φ than itself, and it seems that a system can become or cease to be part of a larger system with higher Φ without making any changes to the system itself.

For example, consider split-brain syndrome. Split-brain syndrome occurs when the connectivity between the left and right hemisphere of the brain has been impaired as a result of a lesion in the corpus callosum (the part of the brain that bridges the hemispheres). Some hold that split-brain syndrome makes consciousness split into two: instead of one unified stream of consciousness whose correlate is distributed across both hemispheres, split-brain patients seem to have two separate streams of consciousness, one for each hemisphere. Theoretically, split-brain syndrome

[5] See Lewis and Langton (1998) for some further qualifications regarding disjunctive or non-natural properties. Also note that intrinsic properties can be subdivided into *absolutely* and *comparatively* intrinsic properties (Pereboom 2015). Comparatively intrinsic are properties that supervene on the extrinsic properties of (or relations between) the *parts* of their bearers. Absolutely intrinsic properties do not supervene on any extrinsic properties. I will use the term intrinsic to include both absolutely and comparatively intrinsic properties. This distinction is especially relevant to Russellian monism, for reasons discussed in footnote 13 below.

[6] By consciousness, I mean *phenomenal* consciousness, the property of there being *something that it is like* for a system to exist, or of having some kind of subjective experience. IIT also defines consciousness in terms of subjective experience, or what it is like to be something (see, e.g., Tononi et al. 2016: 1). By consciousness, I also mean *unified* consciousness, the property of being conscious *as a whole*, as opposed to merely consisting of parts that are individually conscious (as we normally assume is the case for groups of people, or the sum of the two hemispheres of someone with split-brain syndrome) or merely being a part of a larger system that is conscious (as we normally assume is the case for individual neurons in the conscious area of the brain). According to IIT, only *unified* consciousness is restricted to systems with maximal Φ (that is, systems with non-maximal Φ can still be associated with non-unified consciousness by having parts or being part of a whole with maximal Φ and hence unified consciousness).

can also be reversed by healing the lesion. This would also reverse the split in consciousness: at some point during the healing process, the hemispheres will lose their individual streams of consciousness as they "merge" back into single one.

According to IIT, split-brain syndrome occurs because the Φ of some normally maximal brain area distributed across both hemispheres drops to below the Φ its left and right parts as a result of the lesion, and reversal would occur if, by healing the lesion, the distributed area regains maximal Φ (Tononi and Koch 2015: 10). These changes to the corpus callosum need not have any causal effects on the hemispheres themselves in order for maximal Φ to change. Changes in the corpus callosum alone, a system external to both the right and left hemisphere, can thereby be sufficient for them to gain or lose maximal Φ—in the same way a very tall person walking out of or into a room, a change external to every other person in the room, can be sufficient for someone to gain or lose the property of "being the tallest person in the room".

Maximal Φ also depends on circumstances outside the brain. The brain is part of many larger systems that also have some Φ, including the whole body, social groups, nations, the internet,[7] and the solar system. Realistically, the Φ of such systems could never surpass the Φ of the brain, but one might imagine science fiction scenarios in which they do. One such scenario involves grids, i.e., repetitive systems consisting of a potentially very large number of elements, each of which is connected in the same way to the same small number of neighbors. IIT implies that grid-like systems can get very high Φ: if a grid is connected in the right way, its Φ will increase in proportion to the number of elements in the grid, and there is no limit to the number of elements a grid can contain (Tononi et al. 2016; Aaronson 2014).[8]

Now, imagine a set of, say, 10 brains connected in a grid. This grid will have some Φ, but far lower than the Φ of each brain. We then start adding more brains the grid. The Φ of the grid will then increase, but the Φ of each brain and every subsystem within them will remain the same. The Φ of the grid would increase very slowly as more brains are added to it, but in principle, we could keep adding to the grid indefinitely until at some point (perhaps when the trillionth or so brain is added) the Φ of the whole grid will surpass the Φ of each brain and any system within it.[9] The connection

[7] If the internet is understood as a system that includes its users (and their brains), as opposed to just the infrastructure.
[8] This is part of IIT's explanation of why the physical correlate of our consciousness is located where it is within the brain: IIT takes the correlate of our consciousness to be located within the posterior cortex, an area which contains vast grid-like structures and therefore has high Φ (Tononi et al. 2016: 9). The same point figures in Tononi's response to Aaronson's (2014) much discussed objection to IIT, according to which repetitive structures (exemplified by so-called expander graphs) can get very high Φ and therefore be highly conscious according to IIT, which he objects is highly counterintuitive. Tononi's reply (2014) affirms that repetitive structures, including both expanders and grids—which are even simpler and more repetitive than expanders—can be highly conscious, and defends this in part by appeal to how high levels of consciousness seem instantiated by grid-like areas of the brain.
[9] In personal communication, Giulio Tononi has pointed out that if there is any noise or randomness in the connections between the elements of a grid, there will be a point at which adding more elements will no longer increase Φ. In practice, there will always be some noise in a physical structure and this might prevent the possibility of a grid of brains with higher Φ than the brain. But it still seems metaphysically possible for there to be a grid without noise. It would also still be physically possible for the Φ of a grid to surpass the Φ of its elements for other kinds of conscious elements with lower Φ than the brain.

of the final brain will happen far, far away from most of the brains in the network, and it will not causally affect the vast majority of them. But it will still instantaneously make each of them lose maximal Φ—in the same way a tall person walking into the opposite side of a very large room can instantaneously make someone very far away lose the property of being the tallest person in the room.[10]

Maximal Φ is thereby dependent on both close and remote external circumstances and thus clearly fits the description of an extrinsic property. If consciousness is identical to maximal Φ, as IIT claims, it follows that consciousness is also extrinsic and depends on external circumstances in the same way, contrary to our ordinary intuitions.

How could IIT respond to this problem? One response would be to argue that there is no reason to think our pre-theoretical intuitions about consciousness must be reliable and perhaps that it is even to be expected that many of our ordinary intuitions will conflict with our scientific theories. But for those who do not want to abandon intrinsicality,[11] the problem could also be responded to in other ways.

One option is to abandon the *Exclusion* postulate, i.e., the requirement that conscious systems must have *maximal* Φ. Chalmers implies that the *Exclusion* postulate is the only source of the intrinsicality problem, in which case this should be effective—though below I will show that this is actually not the case (it is only the most obvious source). But for now, note that even if abandoning *Exclusion* were sufficient to preserve intrinsicality, this would lead to another highly counterintuitive result, namely that conscious systems massively overlap with other conscious systems. As already noted, the area of the brain that supports our consciousness overlaps with a large number of systems that all have some Φ, including smaller and larger segments of the brain as well as systems that extend far beyond the brain (from social groups to the solar system and beyond). If the *Exclusion* postulate is abandoned, all of these systems will have their own

[10] One might think that in this scenario the brains are not functionally contributing to the grid and thereby not part of it in the sense specified in footnote 2 above, according to which removing the part must affect the functioning of the containing system, because the same grid structure is multiply realizable by other kinds of parts than brains. If the brains are not (functional) parts of the grid, then *Exclusion* would not apply, and they would remain conscious even as the grid surpasses them in Φ. But according to IIT, *Exclusion* applies also in the case of multiple realization, and parts of multiply realizable systems still count as functionally contributing to the containing system because, even though they are replaceable by other kinds parts, removing an actual part *without replacing it* makes a difference to the containing system.

[11] For example, based on what is known as the *Revelation* principle (Strawson 2006a; Goff 2017), according to which phenomenology gives direct insight into the nature of consciousness, combined with the claim that phenomenology directly presents consciousness as intrinsic. However, as will be discussed shortly, given reductionism, intrinsicality implies overlapping consciousness, and it has also been argued that phenomenology presents consciousness as non-overlapping (Goff 2006; though see Chalmers 2016: 190 for criticism). Therefore, appeal to phenomenology does not offer a clear solution the intrinsicality problem for reductionists. Non-reductionism, on the other hand, seems compatible with and arguably even supported by phenomenology and the *Revelation* principle (because phenomenology arguably presents consciousness as irreducible at least to any purely physical constituents (Goff 2017: 147-149)), but as will also be discussed shortly, non-reductionism suffers from other serious problems.

consciousness that overlaps with ours. This may seem at least as counterintuitive as consciousness being extrinsic.[12]

Another way of preserving intrinsicality, but without generating overlapping consciousness, is to reject IIT's identity claim. Clearly, an intrinsic property cannot be identical to an extrinsic property, but they can still be related in other ways. Most obviously, they can be related as cause and effect, because nothing prevents an intrinsic property from being caused by an extrinsic property. The intrinsicality problem could therefore be avoided by rejecting physicalism, the view that consciousness is identical to a physical property, in favor of dualism, the view that consciousness is non-physical but causally related to physical properties. But this solution would not be ideal either, most importantly because dualism faces a serious objection from *the problem of mental causation* (Kim 1989; Papineau 2001). According to this problem, there is good reason to hold that the physical world is causally closed: that every physical event (that has a cause) has a sufficient physical cause—including human actions and other apparently mentally caused events. This suggests that, if consciousness were non-physical, it would be epiphenomenal: unable to causally affect the physical world (except as a redundant over-determiner, a hypothesis that is usually ruled out as *ad hoc*). And epiphenomenalism seems at least as counterintuitive (or otherwise philosophically problematic) as consciousness being either extrinsic or overlapping.

One might think epiphenomenalism could be avoided by combining IIT with Russellian monism rather than dualism. Russellian monism is the view that consciousness is a non-physical property and that either consciousness or protoconsciousness (i.e., properties that are neither physical nor mental, but closely continuous with consciousness) is the categorical[13] ground or realizer of all physical properties, which physics reveals as purely dispositional or relational (Russell 1927; Strawson 2006b). According to its proponents, Russellian monism avoids the main problems of both physicalism and dualism, including the problem of mental causation (Alter and Nagasawa 2012; Chalmers 2013). This is because Russellian monism offers consciousness an explanatory role relative to the physical world, not as an interacting *cause* that generates new causal structure that is not part of physics, but rather as the *realizer* of the same causal structure already posited by

[12] Empirical support for this claim can be found in studies by Knobe and Prinz (2008) showing that people are generally highly resistant to attributing phenomenal consciousness to agents composed of other agents, even if they are otherwise willing to attribute it to systems very different from humans. In support of the validity of such intuitions, Simon (2018) has argued that overlapping consciousness leads to absurd ethical consequences. As previously noted (footnote 11), like intrinsicality, non-overlap has also been supported by appeal to phenomenology.

[13] Russellian monists often claim that phenomenal properties are the *intrinsic* rather than categorical realizers of physical properties, but in the sense of *absolutely* intrinsic rather than *comparatively* intrinsic properties. As noted above (footnote 5), this paper uses the term intrinsic to cover both comparatively and absolutely intrinsic properties (comparatively intrinsic properties supervene on the extrinsic properties of the parts of their bearers; absolutely intrinsic properties do not supervene on extrinsic properties) (Pereboom 2015)). To avoid confusion, I will therefore discuss Russellian monism in terms of categorical properties, which for present purposes can be understood as equivalent to absolutely intrinsic properties.

In terms of absolutely and comparatively intrinsic properties, Russellian monism claims that no physical properties are absolutely intrinsic (though they may be comparatively intrinsic), that merely comparatively intrinsic and extrinsic properties require realizers with absolutely intrinsic properties (or at least some absolutely intrinsic aspects), and that consciousness is absolutely intrinsic (or at least has absolutely intrinsic aspects).

physics. Adding a non-physical realizer to physical causal structure leaves the structure itself unchanged, and so would not imply violation of physical causal closure. But it still gives consciousness an essential explanatory role, assuming that physical structure *requires* some categorical realizers and that there are no physical categorical properties around to do this job.[14]

Russellian monism also goes well with IIT because Russellian monism comes in a panpsychist version (according to which full-blown consciousness rather than mere protoconsciousness realizes all physical structure) and IIT explicitly implies a form a panpsychism (or something very close to it[15]), given that even atoms and protons have some Φ and will therefore have a small amount of consciousness, unless they form part of larger systems with higher Φ (Koch 2012: 132; Tononi and Koch 2015; see also Mørch 2018).[16]

However, unlike dualism, Russellian monism does not necessarily enable a solution to the intrinsicality problem for IIT. This is because the standard version of Russellian monism, known as *constitutive* Russellian monism, takes macroconsciousness (i.e., human and animal-type consciousness) to be constituted by more fundamental microphenomenal[17] (i.e., fundamental particle-type) or protophenomenal properties related in particular ways. If constitutive Russellian monism is combined with IIT and the *Exclusion* postulate, macroconsciousness would be constituted by micro- or protophenomenal properties related in a way that gives rise to maximal Φ. Consciousness would thereby become extrinsic: it would follow that systems would lose or gain consciousness depending on the micro- or protophenomenal properties of external systems.

IIT thereby faces a trilemma between having to deny that consciousness is intrinsic, having to abandon the *Exclusion* postulate and thereby accepting that consciousness massively overlaps, and having to abandon its claim that consciousness is identical (rather than causally or otherwise less intimately related) to maximal Φ (understood as a purely physical property or as including a set of phenomenal or protophenomenal realizers). All of these options are *prima facie* highly

[14] Russellian monism is thereby premised on the rejection of *ontic* structural realism (Ladyman and Ross 2007) and pure dispositionalism (Bird 2007), according to which physical structure does not require any non-structural realizers or physical dispositions do not require any categorical grounds. But it accepts *epistemic* structural realism about the physical world, i.e., that physics only gives us knowledge of the (causal/spatiotemporal/dispositional) structure of the world.

[15] IIT does not imply *complete* panpsychism because simple particles, such as electrons or photons, that may be found in isolation do not clearly have any Φ. But IIT would still seem compatible with the assumption that isolated simple particles have some absolutely simple form of consciousness (or perhaps protoconsciousness, in which case it would also be compatible with panprotopsychism). It is also arguable that simple particles are not actually that simple given quantum field theory and so would actually have some Φ is isolation (Barrett 2014).

[16] It also seems possible to interpret IIT's identity claim as compatible with Russellian monism, insofar as "maximal Φ" can be understood as referring to both a physical property *and* its potentially conscious or protoconscious realizers.

[17] I refer to microphenomenal properties rather than microconsciousness because, as specified in footnote 6 above, I use the term consciousness to imply *unified* consciousness, and given the *Exclusion* postulate, simple systems would lose unified microconsciousness when they come to constitute macroconscious systems, but they could still be regarded as instantiating microphenomenal properties. In general, phenomenal properties can be defined as properties that characterize *what it is like* to be in conscious states, and microphenomenal properties can be defined as properties that characterize what it is like to be in microconscious states (even if these properties may also be experienced from a macroconscious point of view).

counterintuitive—the two former in and of themselves; the latter mainly because it seems to imply epiphenomenalism. But in fact, as I will now show, almost any theory of consciousness faces a trilemma of the same kind. I will then consider IIT's options with respect to the trilemma in more detail.

3 The General Version of the Trilemma

Consider the Global Workspace Theory (Baars 1993; Dehaene and Naccache 2001), one of IIT's most important rivals. This theory claims that consciousness requires implementing a global workspace, which can be very roughly described as a structure that enables information to be accessed and used by several different cognitive subsystems and processes across the brain. But as pointed out by Simon (2018), any area of the brain that constitutes a global workspace will most likely overlap with other smaller or larger areas that also constitute global workspaces. For most global workspaces, it is plausible that if we add or remove one neuron we will still be left with another, slightly smaller global workspace. It follows that consciousness would also massively overlap given the global workspace theory, unless we understand it as implicitly saying that only the *largest* global workspace is conscious, or perhaps only the *most* (i.e., *maximally*) global or "workspace-y" (whatever that might mean) global workspace. But the property of being the largest global workspace requires that the system is not attached to any external components that are also part of a global workspace, and the property of being the most global or "workspace-y" global workspace requires not being attached to any more global or "workspace-y" global workspace, so these properties are extrinsic. It follows that if consciousness is taken as identical to the property of being a (somehow) non-overlapping global workspace, then consciousness is not intrinsic.

We might also consider a radically different theory of consciousness such as the Orch-OR theory (Hameroff and Penrose 1996). The Orch-OR theory claims that consciousness correlates with orchestrated objective reductions, which can be roughly understood as distinctively coordinated (i.e., "orchestrated") sets of quantum collapses happening within microscopic constituents of neurons known as microtubules. Orchestrated objective reductions also seem to overlap because it seems that adding or subtracting a single quantum collapse will usually result in another orchestrated objective reduction. Therefore, the Orch-OR theory must also presuppose that consciousness is correlated with the property of being the *largest*, or perhaps the (somehow) *most* orchestrated, orchestrated objective reduction, which would also be extrinsic.

It thereby seems that most theories of consciousness will face a trilemma between rejecting one of the following claims:

> INTRINSICALITY: Consciousness is intrinsic.
>
> NON-OVERLAP: Conscious systems do not overlap with other conscious systems.
>
> REDUCTIONISM: Consciousness is constituted by more fundamental properties.

REDUCTIONISM is intended to capture both physicalism and constitutive Russellian monism: physicalism (standardly) claims that consciousness is constituted by more fundamental *physical* properties, whereas constitutive Russellian monism claims that macroconsciousness (which the trilemma should be read as being about, given Russellian monism) is constituted by more fundamental *microphenomenal or protophenomenal* properties.[18] Also note that physicalism only implies REDUCTIONISM assuming that consciousness is not to be identified with a fundamental physical property (such as mass, charge or spin). Fundamental physical properties may be intrinsic and non-overlapping; therefore, this kind of physicalism should be excluded from the trilemma. But no current theory suggests that (macro-) consciousness is correlated with a fundamental physical property, so this does not stop the trilemma from applying to most theories of consciousness.

Also note that the trilemma may not hold for theories that take consciousness to be correlated with some *absolute* property such as having some exact value of Φ, being a global workspace of some exact size and so on, because these properties can be both non-overlapping and intrinsic. But any such theory would seem highly implausible, given all the physical differences that exist between individual brains as well as between different conscious states within individual brains.

Another exception would be theories that take consciousness to correlate with a *minimal* property such as implementing the *smallest* global workspace, *smallest* orchestrated objective reduction, and so on. Minimal properties of this sort also do not overlap, and unlike maximal properties they only depend on internal circumstances, i.e., on the (proper) parts of a system *not* implementing a global workspace or *not* being orchestrated, so they are not extrinsic. But theories that correlate consciousness with the smallest systems with some property also seem highly implausible because they could not account for how consciousness typically increases. To illustrate in terms of the Global Workspace Theory, consider a small child (or fetus) whose brain implements the smallest global workspace necessary for consciousness. As the child and its brain grows, so will the child's level of consciousness.[19] But if consciousness depends on the smallest global workspace, the level of the child's consciousness could not increase, because the smallest global workspace cannot grow but only become subsumed by a larger overlapping one (unless the original smallest workspace is changed and restructured during the course of the development so as to no longer constitute a global workspace, but there is no clear reason to think this would necessarily or even typically be the case). Theories that take consciousness to correlate with absolute properties (such as global workspaces of some exact size) would suffer from this problem as well. Absolute and minimal properties seem to be the only intrinsic (and non-fundamental) physical properties that do

[18] Also, given the panpsychist version of Russellian monism, theories of consciousness such as the Global Workspace Theory and Orch-OR should be read as offering a correlate of macroconsciousness, rather than consciousness as such (given that consciousness as such would include fundamental microconsciousness).

[19] By level of consciousness, I mean something like the amount and diversity of information, representations, qualities and other sorts of contents it includes (as opposed to something like intensity of vividness).

not overlap and thereby escape the trilemma. Therefore, the trilemma can be claimed to hold for any *plausible* theory of consciousness.

Now, several aspects of this trilemma have previously been discussed in the philosophical literature. Unger (1980) raises a closely related problem for physical objects known as *the problem of the many*. According to this problem, many physical objects overlap with other objects for the same kind. For example, clouds overlap with a large number of smaller and larger collections of vapor molecules that also seem to qualify as clouds. Unger claims that this gives rise to a dilemma between overlap and nihilism, the latter being the view that clouds and other objects that pose this problem do not exist. Unger also considers whether the problem could be solved by appeal to maximality but concludes that objects such as the largest cloud also overlap. His argument is that the criteria for being a cloud are vague or imprecise, and that the criteria for being the largest cloud are therefore also vague. It follows that there will be many overlapping precise candidates for being the largest cloud.

Unger also considers a version of the problem of the many for consciousness, the *experiential* problem of the many (Unger 2004). He claims that, unlike clouds and other physical objects, consciousness undeniably exists, therefore nihilism is not an option in this case. In order to avoid overlap, he rather proposes abandoning physicalism in favor of a version of substance dualism, according to which a single conscious experience is caused by an overlapping set of nervous systems. For consciousness, he thereby raises a dilemma between non-overlap and reductionism.

The trilemma is similar to the problem of the many applied to consciousness. Like Unger's experiential version of the problem, and unlike the general version, it excludes nihilism as a possible solution (given that the trilemma is essentially a problem for theories of consciousness, which necessarily presuppose that consciousness exists) but rather includes rejecting reductionism. But unlike the experiential version, it also includes the option of appealing to maximality. And, while the general version discusses this option, it does not raise as a problem that it leads to loss of intrinsicality (perhaps because it is less counterintuitive that physical objects such as clouds are extrinsic than that consciousness is). Furthermore, when it comes to the trilemma, appeal to maximality cannot be rejected as ineffective due to vagueness or imprecision, as Unger claims it can for clouds, because the trilemma also holds for theories like IIT that posit a precisely defined correlate of consciousness. Properties such as "being a global workspace" or "orchestration" are not defined equally precisely as Φ, but it is not ruled out that they can be; therefore, the vagueness objection does not necessarily exclude maximality as a solution for these theories either.

A full version of the trilemma has also been raised by Merricks (1998) in an argument against what he calls the doctrine of microphysical supervenience, which is roughly equivalent to REDUCTIONISM. Merricks argues that this reductionist doctrine is false because it contradicts two other claims that he takes to be obvious: first, that consciousness is intrinsic; second, that conscious beings do not overlap with "a mighty host" of other conscious beings. But Merricks' version of the trilemma is not put in terms of properties or systems that may directly correlate with

consciousness in the brain; rather, he argues that *conscious beings* (by which he means whole persons or organisms) have parts that we would also consider conscious in isolation (his main example being some person and the same person minus one finger).[20]

More recently, partial versions of the trilemma have been raised for properties that may directly correlate with consciousness in the brain. Schwitzgebel (2015) defends the conditional claim that *if* physicalism is true, the United States—as well as many other large scale systems that overlap with the brain—are probably conscious. He thereby in effect affirms Unger's dilemma between REDUCTIONISM and NON-OVERLAP for consciousness. His argument is that, given physicalism, consciousness must be a functional property, but all functional properties that could plausibly be identical to consciousness also apply to systems overlapping with the brain. He explicitly takes these functional properties to include brain properties of the sort considered by recent neuroscientific theories. He also considers IIT's appeal to maximality via the *Exclusion* postulate but does not raise the objection that it renders consciousness extrinsic (his main complaint is rather that the postulate seems stipulative and unsupported by argument or evidence). He thereby stops short of identifying the full trilemma.

Simon (2018) has also recently argued that physicalism implies overlapping consciousness, thus also affirming Unger's dilemma between NON-OVERLAP and REDUCTIONISM. As mentioned earlier, Simon also focuses on overlap inside the brain between properties including those characterizing global workspaces and orchestrated objective reductions. He then argues that overlapping consciousness has absurd ethical implications (this complaint is also briefly raised by Merricks (2003)), and that physicalism must therefore be false. Simon also considers whether appeal to maximal Φ offers an escape from the dilemma but, like Schwitzgebel, does not consider the problem that this would render consciousness extrinsic.[21]

To sum up, all plausible theories of consciousness face a trilemma closely related to, but slightly different from, all these previously raised problems: a trilemma between INTRINSICALITY, NON-OVERLAP and REDUCTIONISM. This problem differs from (both versions of) Unger's problem of the many most importantly by (explicitly) including the option of rejecting INTRINSICALITY (although the general version considers the precursor of appealing to maximality) and the fact that this option cannot be rejected due to the vagueness objection. It is close to Merricks' trilemma but differs in that it does not concern conscious beings (i.e., persons or organisms) but rather properties (or

[20] Ted Sider (2003) also affirms this version of the trilemma (but suggests abandoning INTRINSICALITY rather than REDUCTIONISM) in a response to Merricks.

[21] Simon rather objects that maximal Φ actually does overlap because (1) there are multiple candidate mathematical definitions of Φ, and (2) in principle, there can be a tie for maximal Φ if two overlapping systems can have the exact same amount of Φ. But the first objection only seems to show that we do not know which version of Φ (if any) is the one that correlates with consciousness, not that the right version (whatever it turns out to be) would overlap, because the fact that there are multiple candidate versions of a property does not imply that any of the candidates overlap. The second objection is more compelling, and I will consider it in more detail below.

systems) found inside the brain. It also differs from Schwitzgebel and Simon's dilemmas between NON-OVERLAP and REDUCTIONISM by also including INTRINSICALITY.

In view of this trilemma, IIT could respond to the intrinsicality problem by arguing that rejecting INTRINSICALITY is no less plausible than rejecting NON-OVERLAP or REDUCTIONISM, and all theories must reject one.[22] Or, to those who hold that rejecting either NON-OVERLAP or REDUCTIONISM is more plausible, it could be responded that IIT can be rendered compatible with these options as well, by abandoning either the *Exclusion* postulate or the identity claim. I will now consider whether all these options are really compatible with IIT, whether they would really solve the trilemma, and whether they would have any particular advantages or disadvantages for IIT.

4 IIT's Options with Respect to the Trilemma

4.1 Rejecting INTRINSICALITY

Even though preserving INTRINSICALITY contradicts IIT'S identity between consciousness and maximal Φ, rejecting it might seem to contradict IIT in another respect. As mentioned, IIT is supported by a combination of empirical evidence and a philosophical argument based on a set of phenomenological axioms, which are claims about consciousness that IIT takes to be self-evident from reflection on one's own consciousness. These axioms are then "translated" into a set of corresponding physical postulates, according to which conscious physical systems must instantiate structurally similar physical properties. As also mentioned, this argument is highly controversial, but a solution to IIT's intrinsicality problem should still preferably be compatible with it. And it might seem that rejecting INTRINSICALITY is actually not compatible with it, because the first axiom of the argument claims that consciousness "exists intrinsically" (Tononi et al. 2016: 1).

However, the axiom seems to use the term intrinsic in a non-standard sense that does not imply that consciousness is intrinsic in the sense defined here. The first axiom and postulate are described as follows:

> The first axiom of IIT states that experience exists intrinsically. As recognized by Descartes, my own experience is the only thing whose existence is immediately and absolutely evident, and it exists for myself, from my own intrinsic perspective. The corresponding postulate states that the PSC [physical substrate of consciousness] must also exist intrinsically. For something to exist in a physical sense, it must have cause–effect power—that is, it must be possible to make a difference to it (that is, change its state) and it must be able to make a difference to something. Moreover, the PSC must exist intrinsically—that is, it must have cause–effect power for itself, from its own intrinsic perspective. (Tononi et al. 2016: 1)

[22] Above, I showed that each of these options has counterintuitive consequences (and, in footnotes 11 and 12, briefly mentioned some arguments that might support these intuitions), but they may not be equally counterintuitive or otherwise implausible, so this claim should ideally be supported by further arguments than I have mentioned here. In this paper, I will not consider the relative plausibility of the options from the general point of view (as opposed to from the point of view of IIT specifically) in any more detail.

Consciousness is here claimed to exist intrinsically in the sense of existing "for itself". This implies that consciousness has a certain kind of epistemic relation to itself, which according IIT's translation must be reflected physically as having causes and effects upon itself. But necessarily existing "for oneself" does not preclude also necessarily having (or not having) other sorts of relations to other things. Similarly, necessarily having causes and effects upon oneself does not preclude also necessarily having (or not having) causal or other kinds of relations to other things. The axiom can thereby be interpreted not as claiming that consciousness is intrinsic, in the sense of being independent of any external relations between the conscious system and other things, but rather as claiming that consciousness is not *purely* extrinsic, in the sense that, in addition to any external relations, it *also* depends on the conscious system having a particular kind of relation to itself (i.e., the relation of existing for itself, or having causes and effects upon itself). Given this interpretation, there would be no contradiction between the axiom and rejection of INTRINSICALITY.

Still, rejecting INTRINSICALITY might be problematic in other respects, first of all, because IIT might seem to imply a more counterintuitive form of extrinsic dependence than other theories. As discussed above, maximal Φ is a property that depends on both close and remote external circumstances. It is possible to change whether something has maximal Φ by changing circumstances outside, but still directly at the border of, the system, as in the split-brain case, where individual brain hemispheres can gain or lose maximal Φ depending on changes in the corpus callosum, a system which is contiguous with both of them. But it is also possible to change whether something has maximal Φ by changing circumstances far away, as in the grid case, where brains in a grid can lose maximal Φ when the trillionth or so brain is connected to the grid far away from most of the other brains. It might seem more counterintuitive that consciousness depends on such remote external circumstances than on close circumstances directly at their border. In contrast, maximal versions of other properties, such as being the largest global workspace, might seem to only depend on close external circumstances—though this is hard to say, and it might turn out that maximal versions of other properties also depend on remote circumstances in some way.

One might also question the validity of intuition that dependence on remote circumstances is somehow worse than dependence on close external circumstances. The intuition might be due to a confusion between constitutive and causal dependence. Most people find non-local causation (i.e., action at a distance) highly counterintuitive compared to local causation by contact. Non-local constitutive dependence, in contrast, is generally not counterintuitive at all. For example, it is not counterintuitive that the property of being the tallest person in the room depends on the height of people far away in a big room, and it is no less intuitive that it depends of the height of people far away than that it depends on the height of people directly nearby. Someone who accepts that consciousness constitutively depends on close external circumstances, and truly appreciates what

this means (i.e., does not implicitly confuse constitutive with causal dependence), should perhaps therefore not find dependence on remote circumstances any more problematic.[23]

But another problem with rejecting INTRINSICALITY is that it might not preserve NON-OVERLAP in every case, because it is conceivable that two partially overlapping systems have the exact same amount of Φ, resulting in a tie for maximality. This objection is raised by Simon (2018) against the idea that maximal Φ can solve "the problem of the many minds" for physicalism. One response to this problem for IIT is that other theories will face it too: there might be two equally large partially overlapping global workspaces, orchestrated objective reductions, and so on. If this objection holds, the trilemma should perhaps be reduced to a dilemma between NON-OVERLAP and REDUCTIONISM after all. But the tie problem could potentially be solved by stipulating an extra rule for what will happen given ties. For example, one might say that given a tie for maximal Φ, none of the systems involved in the tie will be conscious, but rather only the systems with next to highest Φ.[24] But the problem with such rules is that they might seem highly arbitrary. Another response is to tolerate overlap in the case of ties, because overlap would then be very rare and also involve very few overlapping minds. But those who find overlap objectionable in principle could not accept this.

The tie problem thereby casts some doubt on whether rejecting INTRINSICALITY is a good solution to trilemma for any theory, not just IIT, if the goal is to either preserve NON-OVERLAP specifically or just preserve as many intuitions as possible. This gives additional reason to consider rejecting NON-OVERLAP or REDUCTIONISM instead, for both IIT and other theories.

4.2 Rejecting NON-OVERLAP

NON-OVERLAP is implied by IIT's *Exclusion* postulate, i.e., its claim that consciousness requires maximal, as opposed to merely some non-zero value of, Φ. One problem with abandoning the *Exclusion* postulate is that, unlike INTRINSICALITY (as discussed above), it is clearly intended to follow from a corresponding phenomenological axiom, i.e., the *Exclusion* axiom. But for those who either reject the axiomatic approach or accept it but disagree with this particular axiom (or its translation), it seems the *Exclusion* postulate can be abandoned without great consequences for the rest of IIT.[25]

---

[23] One might think the grid scenario is counterintuitive not mainly because it takes consciousness to depend on remote circumstances but rather mainly because it is highly counterintuitive that grid-like (or repetitive, non-organic-seeming, static, etc.) structures have anything to do with consciousness. This is a valid objection to IIT (influentially raised by Aaronson (2014)), but it is a different objection than the intrinsicality problem, therefore I will set it aside in this paper (as already noted in footnote 1 above).

[24] Assuming there is no tie at this level, too. Could there in principle be a regress of ties all the way down? Given that even protons and neutrons have some Φ, and these entities cannot overlap, it seems not. Even so, it is very counterintuitive that, in principle, a sequence of ties potentially beginning very high up on the macrolevel could be sufficient to reduce consciousness all the way down to the microlevel.

[25] Elsewhere (Mørch 2018), I have pointed out that without *Exclusion*, IIT cannot straightforwardly account for why we ever lose consciousness, given that all brain states (even seemingly unconscious ones such as deep sleep and anesthesia are associated with *some* Φ. Given the *Exclusion* postulate, they could be unconscious because their Φ is

A potential general problem with rejecting the *Exclusion* postulate, however, is that it might seem to give rise to a more counterintuitive kind of overlap than other theories. A property such as being a (non-maximal) global workspace will be instantiated by multiple overlapping segments within the brain. But when it comes to Φ, as already noted, non-zero values are instantiated by systems extending far beyond the brain, from social groups to the solar systems,[26] as well as by microscopic systems all the way down to atoms and protons.

But if overlapping microscopic and extra-cranial consciousness seems more counterintuitive than overlapping, non-microscopic consciousness within the brain, it is probably because the mere *possibility* of consciousness at these levels is counterintuitive in the first place, regardless of whether it overlaps. In other words, the counterintuitiveness primarily derives from IIT's panpsychism. Someone who accepts the possibility of microscopic and extra-cranial consciousness, and also accepts that non-microscopic consciousness overlaps in the brain, should not find it too hard to accept that microscopic and extra-cranial consciousness also overlaps. This objection thereby reduces to the objection that IIT implies panpsychism and panpsychism is counterintuitive, an objection that is distinct from the intrinsicality problem and the trilemma.

Another, more serious problem is that rejecting NON-OVERLAP would actually not succeed in fully preserving INTRINSICALITY. This is because Φ itself (i.e., the absolute, non-maximal version of the property) is not fully intrinsic.[27] To determine the Φ of a given system, one begins by looking at how the present state of a system constrains its possible past and future states: what kind of past states could possibly have caused the present state, and what kinds of future states could the present state possibly cause? One then compares these constraints with the constraints imposed by various partitioned versions of the same system: what kinds of past and future states could possibly have caused and be caused by the same state of each element in the system if the system were cut in two? The greater the difference between the past and future implied by the unpartitioned system and the past and future implied by a partitioned version, the greater the Φ of the unpartitioned system.

But the possible past and future states of a given system (and thus the differences between the possible past and futures of partitioned and unpartitioned systems) do not depend on the present state of the system alone, they also depend on external background conditions that causally

not high enough to be a maximum, but without *Exclusion*, they should be conscious. But IIT could offer alternative explanations, for example, that low-Φ states are only dimly conscious, and therefore not remembered. This explanation may seem somewhat ad hoc (because it is not clear why dimly conscious states cannot be remembered), but it is fully coherent. This point could therefore be regarded as supporting but not necessitating the *Exclusion* postulate.

[26] Though if Schwitzgebel (2015) is correct, all physical properties that can plausibly be identified with consciousness should be understood in abstract, functional terms that imply overlap outside the brain. In that case, extra-cranial overlap would be a problem for everyone, not just explicitly panpsychist theories such as IIT.

[27] Thanks to Giulio Tononi for pointing this out.

influence the system. For this reason, IIT specifies that when calculating Φ, external background conditions should be held fixed:

> Specifically, when evaluating a cause repertoire [i.e., the set of past states that could have caused the present state] in the candidate set [i.e., the system under consideration], the outside elements are fixed at their past state at t_{-1}. Similarly, when evaluating an effect repertoire [i.e., the set of future states that the current state could cause], the outside elements are fixed at their present states at t_0. (Tononi et al. 2014: supplementary text S2, p. 1)

IIT also offers an example of a case where a change in background conditions can change the Φ of a system without changing the system itself:

> Thus, the conceptual structure (C) [i.e., the precise causal structure that determines Φ] of the candidate set may differ, depending on the background conditions, even though t*he state of the elements within the candidate set is the same.*" (Tononi et al. 2014: supplementary text S2, p. 1, my emphasis).[28]

In other words, even though background conditions necessarily (by definition) have the *potential* to causally influence the system, a change in background conditions that does not *actually* causally influence the system can result in a change in its Φ. This shows that Φ itself also satisfies the description of an extrinsic property, or in other words, that it constitutively and not merely causally depends on external circumstances. Abandoning the *Exclusion* postulate, and thus rejecting NON-OVERLAP, is therefore not sufficient to preserve INTRINSICALITY for IIT.

Unlike the problem of ties, according to which rejecting INTRINSICALITY is not sufficient to preserve NON-OVERLAP, this problem does not seem to transfer to other theories: it does not seem (at least not obviously) that non-maximal versions of other properties such as being a global workspace must also depend on background conditions in the same way as Φ does. Those who find overlap acceptable, but extrinsicality and non-reductionism unacceptable, will therefore have reason to reject IIT in favor of another theory. In IIT's defense, one might question why extrinsicality should be regarded as any worse than overlap: aren't these outcomes equally counterintuitive? And why should we trust one of these intuitions more than the other?[29] But

[28] Strictly speaking, a change in the conceptual structure of system (i.e., the specific causal structure of a system that determines its overall Φ) does not imply a change in Φ because different conceptual structures can in principle have the same Φ. According to IIT, a change in the conceptual structure will change the quality but not the level of consciousness, so if the conceptual structure but not Φ were extrinsic the quality of consciousness would be extrinsic (assuming the identity claim) which may be as counterintuitive as consciousness as such being extrinsic. But there is no obvious reason why background conditions should be able to change conceptual structures only in ways that preserve their exact Φ value. In personal communication, Tononi has confirmed the possibility of a change in Φ itself as a result of change in background conditions alone.

[29] According to IIT's argument from axioms to postulates, overlap is definitely worse than extrinsicality, because non-overlap is supported by a phenomenological axiom (i.e., the *Exclusion* axiom which is in turn translated into the *Exclusion* postulate) whereas intrinsicality is not (if my interpretation of the axiom of *Intrinsic Existence*, according to which it is merely intended to characterize consciousness as not purely extrinsic but not thereby intrinsic as defined here, is correct (see previous section)). But it might be objected that, in that case, intrinsicality (as defined here) should be added (or *Exclusion* removed) as a phenomenological axiom.

another possibility would be to consider rather rejecting REDUCTIONISM. I will now argue that, for IIT in particular, this option is more plausible than it may initially appear.

4.3 Rejecting REDUCTIONISM

To recap, the trilemma could also be solved by adopting the dualist view that consciousness is causally produced by purely physical systems with maximal Φ (given IIT) or other extrinsic physical properties (given other theories). This would preserve INTRINSICALITY because nothing prevents intrinsic properties from being caused by, as opposed to identical to, extrinsic properties. But the main problem with dualism is that it arguably leads to epiphenomenalism. The other main non-reductionist view, Russellian monism, arguably avoids epiphenomenalism but standardly comes in a constitutive version according to which, given IIT, macroconsciousness would be identical to maximal Φ *realized by a set of micro- or protophenomenal properties*, which is also an extrinsic property.

But another possibility is to adopt *emergent* Russellian monism. This view takes macroconsciousness to be causally produced by micro- or protophenomenal properties under certain conditions (such as maximal Φ), in accordance with a fundamental law of nature. In other words, it takes macroconsciousness to be a strongly (i.e., nomologically and not merely epistemically) emergent phenomenon. However, emergent Russellian monism tends to be rejected precisely on the basis that, unlike the standard constitutive version of Russellian monism, it fails to avoid dualism's problem of mental causation (Chalmers 2016). Russellian monism offers micro- or protophenomenal properties an explanatory role as the realizer of physical structure, but one might wonder what explanatory role this leaves for macroconsciousness. If macroconsciousness is constituted by micro- or protophenomenal properties related in some way, as per constitutive Russellian monism, it could inherit explanatory relevance from its constituents (analogously to how a macroscopic physical object, such as a hammer, arguably inherits causal relevance from its constituent fundamental particles). But if macroconsciousness is a distinct property causally produced by micro- or protophenomenal properties, as per emergent Russellian monism, it would not inherit their explanatory relevance. Moreover, Russellian monism implies that before brains and other macrophysical systems are formed their physical constituents would already be fully realized by micro- or protoconsciousness. And the formation of the brain and other macrophysical systems does not seem to involve any new, strongly emergent physical structure (according to both the mainstream view of physics, according to which everything macrophysical supervenes on the microphysical, as well as most current theories of consciousness, including IIT). From this, it seems to follow that there would be no physical structure for emergent macroconsciousness to realize, rendering it epiphenomenal.

But there is one version of emergent Russellian monism that avoids this result. According to *the fusion view* of mental combination (Seager 2010; Mørch 2014), when micro- or protoconscious entities come together in the right way, they fuse or "blend" together to form a single unified consciousness. After fusion, the original entities will no longer be conscious or protoconscious on their own: their individual micro- or protoconsciousness will have dissolved into the larger whole.

The new macroconsciousness thereby *replaces* the original micro- or protoconsciousness. Macroconsciousness can therefore take over the explanatory role of the micro- or protoconsciousness it emerged from as the realizer the same physical structure. Before fusion, the particles (or other fundamental constituents) of the (macroconscious parts of) brain were each individually realized by their own microconsciousness, but after fusion, the same particles become jointly realized by a single macroconsciousness instead.

Like standard emergent Russellian monism, the fusion view also takes macroconsciousness to be causally produced by micro- or protoconsciousness under certain conditions, in accordance with a fundamental law, and these conditions may be extrinsic even though the resulting fused consciousness is intrinsic. The fusion view could thereby offer a solution to the trilemma that preserves INTRINSICALITY and NON-OVERLAP but without implying epiphenomenalism. Given IIT, one could posit a fusion law according to which the realizers of physical systems will fuse when a system acquires maximal Φ. The solution is also compatible with other theories. Given the global workspace theory, for example, one could posit a fusion law according to which fusion occurs when systems form a maximal (i.e., largest, most "workspace-y", etc.) global workspace.

But even though the fusion view avoids epiphenomenalism, it has other problems. The main problem with the fusion view compared to other forms of Russellian monism is that it seems inelegant and unparsimonious to posit fundamental fusion laws subsuming the micro- or protophenomenal realizers of some macrophysical properties, assuming (as per the mainstream view of physics) there are no fundamental fusion laws subsuming any macrophysical properties from the physical point of view. However, as I have previously argued (Mørch 2014), it seems fully coherent, even though admittedly somewhat odd and puzzling, that the realizers of a physical system could fuse without this being detectible from the physical point of view.

The main (additional) problem with the fusion view compared to physicalism—the most common way of avoiding epiphenomenalism—is that, as form of Russellian monism, it implies panpsychism or panprotopsychism. Panpsychism is widely seen as highly counterintuitive, and panprotopsychism can be considered obscure and non-explanatory given that we seem to have no independent, positive grasp of the nature of protoconsciousness and how it would be able to explain consciousness. For this reason, someone who is not already committed to Russellian monism (and is also among those who finds pan(proto)psychism counterintuitive) would probably be reluctant to embrace the fusion view as a solution to the trilemma for other theories than IIT: it might seem less plausible to posit both pan(proto)psychism and fundamental fusion laws than to reject INTRINSICALITY or NON-OVERLAP (even granted that it may be more plausible than epiphenomenalism).

But IIT is already committed to panpsychism (or something very close to it), given that, as noted above, even atoms and protons have some Φ. The only problem for this solution for IIT would therefore be that it implies fundamental fusion laws. This is still a significant problem, but perhaps nonetheless preferable to denying INTRINSICALITY or NON-OVERLAP (or accepting

epiphenomenalism). Given this, rejecting REDUCTIONISM specifically in favor of Russellian monism and the fusion view might be the best solution to the intrinsicality problem for IIT.

IIT also goes well with the fusion view for another reason. If the fusion view is correct, it seems plausible that fusion occurs at many levels of reality, rather than just in brains, i.e., that the fundamental fusion laws apply to different kinds of systems that are likely to arise throughout the universe. IIT offers a fusion law of this sort: fusion occurs whenever a system acquires maximal Φ, and maximal Φ can be probably found at all levels of reality and throughout the universe (given that it seems probable that atoms have higher Φ than protons, molecules higher than atoms, cells higher than molecules, and so on). The law that maximal Φ causes fusion is also very simple and elegant. Other theories of consciousness must either offer less general fusion laws (e.g., fusion occurs in global workspaces only) or less simple and elegant ones (e.g., fusion occurs in global workspaces, *and* atoms, *and* molecules, *and* cells, etc.). Therefore, combining IIT with the fusion view may also contribute to somewhat increasing the plausibility of the fusion view.[30]

5 Conclusion

The intrinsicality problem—according to which consciousness and maximal Φ cannot be identical because consciousness intuitively seems intrinsic but maximal Φ is extrinsic—may seem like a serious philosophical problem for IIT. I have shown that the intrinsicality problem actually derives from the fact that, given any plausible theory of consciousness, the following three claims are inconsistent:

> INTRINSICALITY: Consciousness is intrinsic.
>
> NON-OVERLAP: Conscious systems do not overlap with other conscious systems.
>
> REDUCTIONISM: Consciousness is constituted by more fundamental properties.

I have considered whether IIT is really any worse off with respect to this trilemma than other theories. Assuming the solutions are equally plausible, it would not be any disadvantage for IIT if it were committed to rejecting INTRINSICALITY. And, for those who find rejecting INTRINSICALITY not only acceptable but preferable, IIT might even have an advantage, because it identifies

[30] On the other hand, IIT also generates a potential problem for the fusion view. As discussed above, IIT implies that consciousness depends non-locally on remote external circumstances (such as other elements in a large grid). It may follow that the fusion view must take fusion to be non-locally caused. As discussed above, non-local dependence on remote circumstances may seem more counterintuitive than dependence on close circumstances, but this intuition is arguably not valid for constitutive dependence. However, the intuition might still be valid for causal dependence (at least the reasons for dismissing it are less clear).

One might also wonder how the fusion view could deal with the problem of ties. One solution is to stipulate a rule for ties, in the same way reductionists could. But the fusion view could also say that in the case of ties, fusion happens randomly in one of the systems (in the same way that we might think that given a fundamental physical principle of least action, a physical entity might choose a random path in the case of a tie for the most efficient one). Reductionists cannot appeal to randomness in response to ties, because constitution cannot be random.

consciousness with a property that is certain not to overlap,[31] being both determinate and maximal. Many other theories associate consciousness with properties that are not as precisely defined and not explicitly maximal, which could generate overlap between multiple candidate precisifications. Rejecting INTRINSICALITY will therefore more clearly ensure NON-OVERLAP given IIT than given other theories.

For those who find rejecting NON-OVERLAP most plausible, on the other hand, IIT would be at a disadvantage. In IIT's case, rejecting NON-OVERLAP by associating consciousness with Φ *simpliciter* rather maximal Φ will not ensure preservation of INTRINSICALITY, because Φ itself is actually extrinsic, given its dependence on background conditions. But in defense of IIT, one might argue that there is no clear reason to prioritize intuitions against extrinsicality over intuitions against overlap.

For those who find rejecting REDUCTIONISM most plausible, IIT is at no disadvantage. All theories of consciousness, including IIT, are compatible with dualism and emergent Russellian monism, both of which can preserve INTRINSICALITY and NON-OVERLAP by taking macroconsciousness to be causally produced by, rather than identical to, an extrinsic, non-overlapping property. IIT might also have a slight advantage, because it enables a simpler and more elegant version of the emergent Russellian monist fusion view.

Finally, for those who find IIT plausible in itself, but are neutral between the different solutions to the trilemma, the solution of rejecting REDUCTIONISM in favor of the Russellian monist fusion view should seem especially plausible. This solution implies pan(proto)psychism, which many find counterintuitive, but IIT already implies (something close to) panpsychism anyway. The main problem with this solution for IIT is that it posits fundamental fusion laws, but this might seem preferable to having to deny INTRINSICALITY or NON-OVERLAP (or accept epiphenomenalism).

To sum up, the intrinsicality problem for IIT is not as serious as it may seem. In view of the trilemma, it is far more defensible to deny that consciousness is intrinsic. IIT is also compatible with solutions that need not deny that consciousness is intrinsic, although the set of such solutions is more limited for IIT than for some other theories: it includes solutions that reject REDUCTIONISM but excludes solutions that reject NON-OVERLAP. But this is not necessarily a problem given that, for IIT, rejecting REDUCTIONISM in favor of the Russellian monist fusion view might be the most plausible solution anyway.

Acknowledgements
Many thanks to Jessie Munton and Giulio Tononi for comments and discussion.

[31] Except for the problem of ties.

References

Aaronson, Scott. 2014. Why I Am Not an Integrated Information Theorist (or, the Unconscious Expander). *Shtetl-Optimized: The Blog of Scott Aaronson*, https://www.scottaaronson.com/blog/?p=1799.

Alter, Torin, and Yujin Nagasawa. 2012. What Is Russellian Monism? *Journal of Consciousness Studies* 19 (9-10): 67-95.

Baars, Bernard J. 1993. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Barrett, Adam B. 2014. An Integration of Integrated Information Theory with Fundamental Physics. *Frontiers in Psychology* 5: 63.

Bayne, Tim. 2018. On the Axiomatic Foundations of the Integrated Information Theory of Consciousness. *Neuroscience of Consciousness* 2018 (1): niy007-niy007.

Bird, Alexander. 2007. *Nature's Metaphysics: Laws and Properties*. Oxford: Clarendon Press.

Cerullo, Michael A. 2015. The Problem with Phi: A Critique of Integrated Information Theory. *PLOS Computational Biology* 11 (9).

Chalmers, David J. 2013. Panpsychism and Panprotopsychism. *The Amherst Lecture in Philosophy* 8 (1-35.): Reprinted in Brüntrup and Jaskolla 2016.

Chalmers, David J. 2016. The Combination Problem for Panpsychism. In *Panpsychism: Contemporary Perspectives*, eds. G. Brüntrup and L. Jaskolla. Oxford: Oxford University Press.

Dehaene, Stanislas, and Lionel Naccache. 2001. Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition* 79 (1-2): 1-37.

Goff, Philip. 2006. Experiences Don't Sum. *Journal of Consciousness Studies* 13 (10-11): 53-61.

Goff, Philip. 2017. *Consciousness and Fundamental Reality*. Oxford University Press.

Hameroff, Stuart, and Roger Penrose. 1996. Orchestrated Objective Reduction of Quantum Coherence in Brain Microtubules: The" Orch or" Model for Consciousness. *Mathematics and computer simulation* 40: 453-480.

Kim, Jaegwon. 1989. Mechanism, Purpose, and Explanatory Exclusion. *Philosophical Perspectives* 3: 77-108.

Knobe, Joshua, and Jesse J. Prinz. 2008. Intuitions About Consciousness: Experimental Studies. *Phenomenology and the Cognitive Sciences* 7 (1): 67-83.

Koch, Christof. 2012. *Consciousness: Confessions of a Romantic Reductionist*. Cambridge, MA: MIT Press.

Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Clarendon Press.

Lewis, David, and Rae Langton. 1998. Defining 'Intrinsic'. *Philosophy and Phenomenological Research* 58 (2): 333-345.

Merricks, Trenton. 1998. Against the Doctrine of Microphysical Supervenience. *Mind* 107 (425): 59-71.

Merricks, Trenton. 2003. Maximality and Consciousness. *Philosophy and Phenomenological Research* 66 (1): 150-158.

Mindt, Garrett. 2017. The Problem with the 'Information' in Integrated Information Theory. *Journal of Consciousness Studies* 24 (7-8): 130-154.

Mørch, Hedda Hassel. 2014. Panpsychism and Causation: A New Argument and a Solution to the Combination Problem (Doctoral Dissertation), Departement of Philosophy, Classics, History of Art and Ideas, University of Oslo, Oslo.

Mørch, Hedda Hassel. 2018. Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? *Erkenntnis*.

Papineau, David. 2001. The Rise of Physicalism. In *Physicalism and Its Discontents*, eds. C. Gillett and B. Loewer. Cambridge: Cambridge University Press.

Pereboom, Derk. 2015. Consciousness, Physicalism, and Absolutely Intrinsic Properties. In *Consciousness in the Physical World: Perspectives on Russellian Monism*, eds. T. Alter and Y. Nagasawa Oxford University Press.

Russell, Bertrand. 1927. *The Analysis of Matter*. London: Kegan Paul, Trench, Trubner & Co.

Schwitzgebel, Eric. 2015. If Materialism Is True, the United States Is Probably Conscious. *Philosophical Studies* 172 (7): 1697-1721.

Seager, William. 2010. Panpsychism, Aggregation and Combinatorial Infusion. *Mind and Matter* 8 (2): 167-184.

Sider, Theodore. 2003. Maximality and Microphysical Supervenience. *Philosophy and Phenomenological Research* 66 (1): 139-149.

Simon, Jonathan A. 2018. The Hard Problem of the Many. *Philosophical Perspectives*.

Strawson, Galen. 2006a. Panpsychism? Reply to Commentators with a Celebration of Descartes. *Journal of Consciousness Studies* 13 (10-11): 184-280.

Strawson, Galen. 2006b. Realistic Monism: Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies* 13 (10-11): 3-31.

Tononi, Giulio. 2014. Why Scott Should Stare at a Blank Wall and Reconsider (or, the Conscious Grid). *Shtetl-Optimized: The Blog of Scott Aaronson*, https://www.scottaaronson.com/tononi.docx.

Tononi, Giulio, Larissa Albantakis, and Masafumi Oizumi. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology* 10 (5).

Tononi, Giulio, Melanie Boly, Marcello Massimini, and Christof Koch. 2016. Integrated Information Theory: From Consciousness to Its Physical Substrate. *Nature Reviews Neuroscience* 17 (7): 450-461.

Tononi, Giulio, and Christof Koch. 2015. Consciousness: Here, There and Everywhere? *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 370 (1668): 20140167.

Unger, Peter. 1980. The Problem of the Many. *Midwest Studies in Philosophy* 5 (1): 411-468.

Unger, Peter. 2004. The Mental Problems of the Many. In *Oxford Studies in Metaphysics, Vol. 1*, ed. D. Zimmerman Oxford: Clarendon Press.