***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

Integrated Information Theory, Intrinsicality, and Overlapping Conscious Systems

James C. Blackmon

Abstract

Integrated Information Theory (IIT) identifies consciousness with having a maximum amount of integrated information. But a thing's having the maximum amount of anything cannot be intrinsic to it, for that depends on how that thing compares to certain other things. IIT's consciousness, then, is not intrinsic. A mereological argument elaborates this consequence: IIT implies that one physical system can be conscious while a physical duplicate of it is not conscious. Thus, by a common and reasonable conception of intrinsicality, IIT's consciousness is not intrinsic. It is then argued that to avoid the implication that consciousness is not intrinsic, IIT must abandon its Exclusion Postulate, which prohibits overlapping conscious systems. Indeed, theories of consciousness that attribute consciousness to physical systems, should embrace the view that some conscious systems overlap. A discussion of the admittedly counterintuitive nature of this solution, along with some medical and neuroscientific realities that would seem to support it, is included.

1. Introduction

Integrated Information Theory (IIT), originally developed by Giulio Tononi (2004), is a theory of consciousness that identifies the property of having consciousness with the property of having a "maximum of integrated information" (Φ^{max}), the detection of which is held to be achievable by empirically examining the causal structure of a physical system of elements (e.g., neurons or logic gates).[1] Criticism of IIT has tended to involve objections that IIT attributes consciousness to physical systems which we should regard as not conscious or fails to attribute consciousness to systems we should regard as conscious.[2] The argument here is that IIT in its current form implies that consciousness is not intrinsic according to a common and reasonable account of intrinsicality. It will then be argued that to avoid the implication that consciousness is extrinsic, IIT must reject its prohibition on overlapping conscious systems, found in its Exclusion Postulate. So as to preserve the intrinsicality of consciousness, a theory of consciousness in physical systems must embrace overlap: the brain hosts many conscious entities at once, many of them overlapping each other—an undoubtedly strange thesis that surely demands, and will receive here, further support.

This is not the first argument that finds a challenge for IIT in the apparent intrinsicality of consciousness. John Searle (2013) argues that the theory's foundation in information falsely implies consciousness is an observer-relative, and thus extrinsic, property.[3] David Chalmers (2016, fn 8) briefly notes that IIT seems to entail that consciousness is extrinsic. Hedda Hassel Mørch (2019) explores this point in depth, finding not a dilemma between rejecting intrinsicality and embracing overlap, but an inconsistent triad:

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

[1] Tononi (2004), Tononi (2015), Oizumi *et al.* (2014), Tononi et al. (2016)
[2] Searle (2013), Aaronson (2015), Schwitzgebel (2015), and Horgan (2015)
[3] See Koch and Tononi (2013) and Fallon (2020) for discussion of Searle's view.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

Mørch argues that IIT can and should reject REDUCTIONISM so as to preserve INTRINSICALITY and NON-OVERLAP on the basis that doing so is the least counterintuitive option. Mørch defines intrinsicality as follows:

> 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.

While Mørch does not clarify the concept of constitutive dependence, it becomes clear that rejecting REDUCTION does not amount to rejecting all forms of acausal dependence on external conditions. Indeed, as we will see, even if consciousness does not constitutively depend on external conditions, whether IIT considers a system to be conscious can nevertheless depend on acausal relations, i.e., mere comparisons, to external factors. Acausal independence of any kind, not just constitutive independence, would seem to better capture our notion of intrinsicality as it would apply to consciousness.

As we will use the term here, intrinsicality is construed in a Lewisian spirit to meet the following necessary condition:

> If property *P*, holding of some physical system *x* is intrinsic to *x*, then any physical duplicate of *x* also has *P*.

Your nonrelativistic mass, blood type, and body temperature are intrinsic to you because a physical duplicate of you must have the same non-relativistic mass, blood type, and body temperature. Your weight, location, and access to Wi-Fi are not intrinsic to you because a physical duplicate of you can vary in these regards. Note that no commitment to a property's constitution or constitutive dependence is made here. On this analysis, to say that consciousness is intrinsic is not simply to rule out constitutive dependence on external conditions; it instead rules out dependence on any acausal external conditions.

We will proceed using this particular conception of intrinsicality, but the general point can be made without it. As mentioned, IIT identifies consciousness as the property of having a "maximum of integrated information", Φ^{max}, a property which holds for a physical system if and only if that system has more of a certain kind of causal integration than any other physical systems overlapping it.[4] By this definition, Φ^{max} is clearly an extrinsic property: whether a system has Φ^{max} depends on how much of this kind of causal integration other overlapping physical systems have. Indeed, having the maximum amount of anything (being the tallest, being the oldest) depends strictly on acausal relations, mere comparisons. As such, Φ^{max} is like being the most attractive person at the party, a property that can come and go, as do the guests, without entailing any relevant physical change in the person who gains or loses the property. But if consciousness and Φ^{max} are identical, then having consciousness depends on mere comparisons, specifically, whether certain comparative and acausal relations to certain other physical systems hold.[5]

It will be argued that IIT can and should abandon the identification of consciousness with Φ^{max} and admit that overlapping physical systems can be independently conscious. On this alternative, physical systems can

[4] Integrated information (Φ) "quantifies to what extent the cause-effect structure specified by a system's elements changes if the system is partitioned (cut or reduced) along its minimum partition (the one that makes the least difference)." Integrated information is a measure of the causal interdependence of a system's elements, where a $\Phi=0$ indicates that every element functions independently of every other part.

[5] As such, the consciousness of physical systems would seem to "violate causality"; that is, distant conditions would instantly determine whether that system is conscious. For a related point, see van Stekelenburg and Edwards (2020).

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

still have Φ^{max}, and such systems can still be conscious by some related measure, but other overlapping systems can be conscious as well.

There is another source of extrinsicality in IIT. As Mørch notes, Tononi accepts that the amount of Φ in a system depends on background conditions which could be different. Φ, in this sense, is like potential energy. Consequently, rejecting the Exclusion Postulate is an insufficient, though necessary, condition for preserving intrinsicality. This matter will be discussed briefly in the conclusion.

2. Exclusion

IIT's prohibition on overlapping conscious systems is made explicit in the first clause of its Exclusion Postulate.

> Exclusion: Of all overlapping sets of elements, only one set can be conscious – the one whose mechanisms specify a conceptual structure that is maximally irreducible (MICS) to independent components. A local maximum of integrated information Φ^{Max} (over elements, space, and time) is called a complex. (Oizumi et al. 2014)[6]

The prohibition on overlap has been explained in other terms elsewhere. Tononi *et al.* (2016) address Φ^{max} independently of its postulated identity as consciousness.

> Because a prerequisite for intrinsic existence is having irreducible cause-effect power, the cause-effect structure that actually exists, over a set of elements and spatio-temporal grains, is that which is maximally irreducible (Φ^{max}), called a conceptual structure. As a consequence, any cause-effect structure overlapping over (*sic*) the same set of elements and spatio-temporal grain is excluded.

Referring to what is termed the *physical substrate of consciousness* (*PSC*), the authors make room for coexisting complexes in a single brain so long as they do not overlap, writing, "According to IIT, two or more non-overlapping complexes may coexist as discrete PSCs within a single brain, each with its own definite borders and value of Φ^{max}." Tononi (2015) writes, "if there are alternative assignments of cause-effect repertoires (purviews) over subsets of elements within a complex, the winner is the assignment that supports the conceptual structure having Φ^{max}."

In what follows, we will see that the dictum "There can be only one", that is, the insistence that there be some anointed "winner", is what implies that consciousness depends on mere comparison.

3. A Mereological Placement Argument

The argument presented here is intended to show that physical systems can differ in whether they have Φ^{max} even when they are physical duplicates. It follows that Φ^{max} is not intrinsic.

> Let S and T be physical systems such that S is a proper part of T, S is causally connected to the rest of T through connections c_i, and T has Φ^{max}. T might be some region of a human brain with Φ^{max}. S would then be some proper part of that brain region. Or T might be Tononi's (2015) "didactic

[6] The Exclusion Postulate does not address what happens if there are two or more systems with maximal Φ. Mørch calls this *the problem of ties*.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

> example", the complex **ABC** (whose elements **A**, **B**, and **C** are logic gates), which has Φ^{max}. S could then be any proper part, for instance, **AB**. The connections c_i would be just like the connections between the units such as **A** and **B**. Because S and T share parts, S and T overlap.[7] Thus, by the Exclusion Postulate, S is not conscious because it overlaps T which, having Φ^{max}, is conscious. IIT makes this much explicit.
>
> Now, let S* be physical duplicate of S, but instead of being physically connected to an enclosing system in the way that S is connected to the rest of T, let S* be connected through connections c_i* which causally affect S* in exactly the way the c_i causally affect S.[8] The preceding establishes that, although S and S* are physical duplicates, and although they are causally affected in identical ways, they are embedded in physically different systems. Finally, let S* have no overlapping parts with more Φ than S* has. S*, then, has Φ^{max}, and thus, according to IIT, S* is conscious. But S is not conscious, even though S and S* are physically identical. Therefore, consciousness, under IIT, is not an intrinsic property. Whether a physical system has it depends on acausal relations to other things, mere comparisons.

The placement argument shows that unless consciousness is not an intrinsic property as we use the term here, consciousness is not Φ^{max}. In fact, the placement argument appears to provide this result using other accounts of intrinsicality. Consider other pre-theoretic accounts given by Lewis (1983).

> A sentence or statement or proposition that ascribes intrinsic properties to something is entirely about that thing; whereas an ascription of extrinsic properties to something is not entirely about that thing, though it may well be about some larger whole which includes that thing as part.
>
> A thing has its intrinsic properties in virtue of the way that thing itself, and nothing else, is.

The placement argument also shows that consciousness does not supervene in certain ways on the physical or microphysical properties of its instantiation. Consider, for instance, Kim (1987).

> *A*-properties *weakly supervene* on *B*-properties if and only if for any possible world *w* and any individuals *x* and *y* in *w*, if *x* and *y* are *B*-indiscernible in *w*, then they are *A*-indiscernible in *w*.

S and S* are physically indiscernible, though on IIT they differ regarding consciousness.

Note that these consequences follow whether we regard consciousness as reducible to more fundamental properties. For example, if IIT adopts epiphenomenalism, then consciousness is caused, not constituted, by Φ^{max}, but as the placement argument shows, S still is not conscious while S* is, showing that consciousness is still not an intrinsic property. Thus, while rejecting Mørch's REDUCTIONISM does avoid the consequence that IIT's consciousness is constitutively dependent on external conditions, doing so does not avoid the consequence that IIT's consciousness is not intrinsic in the Lewisian sense we use here or in other senses, nor does it avoid the consequence that consciousness does not supervene in the sense described. IIT's consciousness continues to depend on external conditions in acausal ways that do not involve constitution.[9]

---

[7] *x* overlaps *y* if and only if *x* and *y* share parts. Because these shared parts need not be proper parts of *x* and *y*, S overlaps T when S is a proper part of T. See Casati and Varzi (1999).

[8] The c_i* can be utterly simple; they need only send the right signal at the right time. They can achieve this in complete independence of each other and each with nothing more than a fixed track of signal commands set for the right times.

[9] Even if IIT retreats from the identity of consciousness with Φ^{max} to the view that Φ^{max} is a necessary condition for consciousness, this does not reclaim intrinsicality, for the placement argument shows that whether a thing meeting S's

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

To appreciate the costs of IIT's Exclusion Postulate, consider this brief survey of some of its controversial implications in light of the placement argument, implications which appear to hold regardless of whether REDUCTIONISM is rejected.

First, the Exclusion Postulate entails that philosopher's zombies can be, and likely are, a physical reality in the actual world. Oizumi *et al.* (2014) have already shown that IIT permits "zombies" (their quotes) of a kind: systems which are input-output equivalent to conscious systems but not integrated enough to be conscious themselves. However, the authors stop short of admitting philosopher's zombies as traditionally understood. The philosopher's zombie is a hypothetical being defined as being physically identical to a conscious being in every way but lacking consciousness. They are typically proposed to be merely conceivable by us, not real and not even possible in any stronger sense. The point of invoking these zombies is not to raise the concern that it is physically possible to have them among us in the actual world, concerning as that might be; the point is instead to argue that their conceivability somehow demonstrates that one's consciousness does not follow directly from one's complete physical description. When one wishes to defend the reality of the Explanatory Gap or the Hard Problem, or to advance a form of dualism, one musters the zombies, and it is their conceivability alone that constitutes their force. But as we have just seen, S* is conscious, while physically identical S is not, and both S* and S are physically possible in this world; thus, S is a physically possible zombie in this world. Given that IIT attributes consciousness to relatively simple systems, and given the apparent likelihood that some of these simple systems have duplicates embedded in systems with Φ^{max}, IIT appears committed to the actual existence of zombies in our world. Moreover, you could be transformed into such a zombie, according to IIT's Exclusion Postulate, by brilliant, powerful, and unkind neuroscientists who manage to embed your PSC in a larger system which has Φ^{max} so that your PSC evolves precisely as it would have had you been left alone. They might instead alter the system your PSC is already embedded in so that this embedding system has Φ^{max} but your PSC evolves just as it would have had you been left alone. In cases like these, your PSC is initially like S* until these neuroscientists bring it about that your PSC is like S, embedded in a larger T-like system which has your PSC as a proper part and which lays exclusive claim to Φ^{max}. According to the Exclusion Postulate, your PSC is no longer conscious, although it is physically identical to the PSC that would have been conscious had this operation never been performed.[10]

Second, in these embedding cases, and assuming that one can know that one is conscious by introspection, one can go from knowing that one is conscious to not knowing whether one is conscious (perhaps not knowing anything, depending on whether knowledge requires consciousness) even though one's PSC is physically indistinguishable from how it would have been had it never been embedded—in which case one could continue to be conscious and to know it. Knowledge of one's consciousness is extrinsic just as consciousness is, according to IIT.

Third, and still assuming that one can know that one is conscious by introspection, one's knowing that one is conscious would be evidence (whether one recognizes it as such or not) that one is not embedded in a system with Φ^{max}; it would be evidence about what kinds of things do and do not furnish one's external world. A system such as S* that is conscious and knows it can infer correctly that it is not embedded in a system with Φ^{max}. If you know by introspection that you are conscious, then you can infer, without further empirical

physical description (thereby meeting S*'s physical description) could possibly be conscious still depends on mere comparisons.

[10] Horgan (2015) makes the similar but different observation that a system of conscious humans would become zombies if they ever interacted in such a way as to create a system with more Φ. In Horgan's example, the zombification occurs only once the humans behave in intrinsically different ways, whereas S and S* behave exactly the same.

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

investigation, that no region that incorporates you, in the entire universe, has Φ^{max}. You are the "winner" among all these uncounted regions of the universe, however variegated or vast, and to the extent that you know IIT is true, you know it.

Fourth, and still assuming that one can know that one is conscious, S* has S as its non-conscious physical duplicate, so S would be going through the motions of knowing one is conscious but inferring (or "inferring" depending on whether such inferences require consciousness) incorrectly that it is not embedded in a system with Φ^{max}. Knowledge of consciousness loses all physical relevance in these cases.

IIT's Exclusion Postulate entails a world in which non-humanoid zombies almost certainly exist, actual human zombification is physically possible, robust empirical knowledge about the external world can be gained by introspection, and knowledge of one's consciousness is significantly irrelevant. These consequences follow whether consciousness constitutively depends on external features or just acausally depends on them in some other way. Accepting IIT in its current form comes at the price of radically revising our understanding not only of consciousness but of our epistemic relation to the empirical world. Unless we have good independent reasons for accepting these results, the burden of explanation appears to be on the advocates of IIT.

4. Embracing Overlap

IIT, and any theory of consciousness for physical systems, should permit overlapping conscious systems. This way, both S and S* can be conscious to some degree determined only by their intrinsic physical properties and not merely by their placement among other things.

Granted, many people will oppose the suggestion that conscious systems can overlap. There are at least two intuitive reasons why one might do so. I suggest here that both reasons, while alluring, fail to justify the overlap prohibition.

First, one might appeal, as Tononi does, to parsimony.

> "The exclusion postulate can be said to enforce Occam's razor (entities should not be multiplied beyond necessity): it is more parsimonious to postulate the existence of a single cause-effect structure over a system of elements—the one that is maximally irreducible from the system's intrinsic perspective—than a multitude of overlapping cause-effect structures whose existence would make no further difference." (Tononi 2015)

In prohibiting this multitude of overlapping cause-effect structures, a multitude of overlapping conscious entities is also prohibited. IIT is thus relatively parsimonious regarding the *number* of things in the universe that are conscious.[11]

But IIT is not so parsimonious regarding the *kinds* of physically identical things that exist, nor is it as parsimonious regarding the *number of things on which consciousness depends*. As the placement argument shows, for any physical description of an object with any non-zero level of integrated information, there are two possible kinds of things: the unconscious S-like kind and the conscious S*-like kind. Moreover, as the conditions

[11] IIT is not especially parsimonious in this way since most competing theories do not regard every physical system with Φ^{max} to be conscious and are thus significantly more parsimonious than IIT. IIT is, however, more parsimonious in this way than the overlap alternative we are considering.

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

surrounding such an entity change in any way that alters whether it has Φ^{max}, so does that entity's status as a conscious or unconscious entity. On IIT, consciousness springs into and out of existence, sometimes due only to causally irrelevant changes around, not within, physical substrata. IIT's ontology is in this way not especially parsimonious.

On the other hand, if overlap is permitted, we may have only different overlapping physical systems with their degrees of integrated information and their corresponding conscious states. To be sure, the overlap alternative recognizes more conscious entities in the universe, but by simpler means.[12] On the overlap alternative, both S and S* may be the same kind of thing regarding consciousness (both conscious or both not conscious), and whether they are conscious depends only on their intrinsic nature, not on causally irrelevant differences existing elsewhere.

A second reason for rejecting overlapping conscious entities is that it can just seem obviously wrong or absurd, perhaps due to how we understand consciousness by way of introspection: "I obviously do not experience overlapping conscious entities, nor do I ever have other experiences along with whatever present experience I might be having. In short, there's nobody here but me!"

We should grant that the idea of having many overlapping conscious entities in one's head (yes, the irony of "*one*" is noted) can at first seem unacceptable, obviously false, counterintuitive, or even disturbing. In fact, this denial of Cartesian indivisibility is counterintuitive not only because we do not expect conscious systems to overlap each other but also because we do not expect a plurality of them in ourselves and in other people, nor do we expect conscious systems to be composed of other conscious systems.[13] But the overlap view, as expressed here, includes them all. Readers who balk at overlap are not without esteemed company.

Hilary Putnam's (1975) early functionalism contains an ostensibly *ad hoc* rider intended to rule out cases of overlap such as swarms of bees as "single pain-feelers". Putnam gives no further comment, suggesting that he felt it was simply obvious that conscious entities could not overlap.

Ned Block's (1978) "Chinese Nation" thought experiment in which the people of China come to realize the functional structure of pain is considered by Block to be among his "troubles" with functionalist accounts of qualia.

Arguing against a form of supervenience, Trenton Merricks (2003) rejects overlapping conscious entities, asserting that "there is exactly one conscious being—me—now wearing my shirt, now sitting in my chair." He elaborates by noting that if many conscious beings were sitting in his chair and thinking his thoughts, then some would survive a haircut while others would not and his marriage would amount to polygamy. For Merricks, inadvertent polygamy and death by haircut form a *reductio* sufficient for dismissing the prospect of overlapping conscious beings.[14] From this conviction, Merricks rejects Microphysical Supervenience.

---

[12] Recognizing proper parts of things already recognized does not "pack the universe fuller", even if it produces a "longer list" of things existing, nor are we eating less cake by refusing to acknowledge the existence of the its left and right halves.

[13] Mørch cites Knobe and Prinz (2008) as showing just how averse we are to conceiving of conscious beings as being composed of conscious beings. In the absence of a cogent argument, the intuition's stubbornness should make us all the more suspicious of it. Perhaps we are merely in the grip of a tenacious cognitive bias.

[14] As Mørch points out, Merricks is not predicating consciousness of systems or brain regions. He is speaking of conscious beings such as the being composed of all of a person's atoms but for one atom in that person's left index finger.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

Peter Unger (2004) is scandalized by the notion of overlapping conscious entities, writing, "The thought that there are, in my situation, vastly many individuals each similarly experiencing the sweet taste of chocolate is, to my mind, a very disturbing suggestion." Unger is clear that he does not regard this as merely another version of his original Problem of the Many (Unger 1980) according to which there is a puzzling semantic mystery whether what we regard as a cloud (or a chair, or a human) is actually many overlapping clouds (or chairs, or humans) sharing many but not all parts. Instead, he finds what he dubs "the problem of many experiencers" and the "problem of many choosers" to be deeply disturbing for its metaphysical implications. His solution is to adopt substance dualism.

Searle (2013), too, appears to take it as obvious that conscious entities cannot overlap when he uses this premise in an attempted *reductio* of both IIT and panpsychism.

> Consciousness comes in units. The qualitative state of drinking beer is different from finding the money in your wallet to pay for it. But a consequence of its subjectivity is its unity. So for example, I am conscious and you are conscious but each consciousness is separate from the other; they do not smear into each other like adjoining puddles of mud. Consciousness cannot be spread over the universe like a thin veneer of jam; there has to be a point where my consciousness ends and yours begins. For people who accept panpsychism, who attribute consciousness, as Koch does, to the iPhone, the question is: Why the iPhone? Why not each part of it? Each microprocessor? Why not each molecule? Why not the whole communication system of which the iPhone is a part? The problem with panpsychism is not that it is false; it does not get up to the level of being false. It is strictly speaking meaningless because no clear notion has been given to the claim. Consciousness comes in units and panpsychism cannot specify the units.

Putnam rules out the possibility of overlapping conscious entities by stipulation, and Block counts it a trouble. Merricks finds that it leads to absurdities regarding the possession of shirts and spouses. Unger expresses utter dismay at the possibility. Searle makes bald assertions which deny it. But a substantive argument that isn't haunted by the ghost of Cartesian indivisibility is hard to find. Instead, it appears that the main reason for rejecting the possibility of overlapping conscious entities is that it is simply unpalatable.[15]

Fortunately, palatability is not our only guide, and in this case, it threatens to disguise a fallacy. Statements about what I do not experience presuppose a particular experiencer, an entity which can represent itself and refer to itself using singular first-person pronouns. Thus, there exists an *x* with the conscious cognitive and causal capacities to report that *x* does not experience any overlapping conscious entities. But this would disprove the existence of other overlapping conscious entities no more than it would disprove the existence of other conscious entities anywhere at all. Taking an introspected lack of evidence as introspected evidence of lack is an argument from ignorance—a simple (if pardonable) fallacy. For all we know, each introspecting brain may host many "introspectors", each partaking in similar introspective activities yielding a chorus of reports: "Nobody here but me!"

Furthermore, why would a conscious entity necessarily be aware of another conscious entity overlapping it, especially if they work in relative harmony? If overlapping conscious entities reside in humans, they obviously

[15] Mørch, too, finds OVERLAP to be counterintuitive, and I think we must agree that initially it is. Importantly, while Mørch does find OVERLAP's counterintuitive nature to weigh against it, she does not insist on the strength of its seeming implausibility that it is clearly false or unacceptable. Sider (2003), who distinguishes between consciousness and consciousness* and Schwitzgebel (2020) appear to be more willing to accept overlapping conscious regions. Sider's consciousness* appears equivalent to the concept of consciousness used here.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

must be causally integrated enough to be able to speak and write, catch fly balls, and descend crowded staircases while carrying hot coffee and scrolling the news feed on one's phone. They will not, in most cases, regard each other as Others in any obvious representational, *de dicto*, or propositional sense, as external beings with which they must communicate, collaborate, or compete.[16] Indeed, it may be that for some slightly overlapping conscious systems *x* and *y*, *x* largely receives the products of *y*'s conscious processing (and *vice versa*), as perhaps thoughts just seem to "come to" *x*. Moreover, if *x* and *y* overlap sufficiently, most of their conscious states will be so similar that these states would be difficult for *x* and *y* to distinguish were they ever somehow given the chance. But even once *x* has a reason to suspect there is such a *y*, *x* cannot be expected to discover *y* through normal introspection; after all, introspection will typically be carried out not by one conscious entity attempting to discover others by observing itself, but by numerous overlapping and causally intertwined conscious entities attempting to discover others by observing themselves.

Whether any advocates of IIT use the introspected lack of evidence as if it is an introspected evidence of lack is not clear, but I do think the tendency to find singularity in introspection is there for many of us even if we have come to repudiate it. Regarding a particular conscious experience of one's hands on a piano, Tononi *et al.* (2016) write,

> "the content of my present experience includes seeing my hands on the piano, the books on the piano, one of which is blue, and so on, but I am not having an experience with less content (for example, the same scene in black and white, lacking the phenomenal distinction between coloured and not coloured) or with more content (for example, including the additional phenomenal distinction of feeling one's blood pressure as high or low)."

Tononi's description presupposes—does not establish—a sole experiencer. We may grant that there is a conscious entity which experiences seeing the hands, books, and so on, and moreover that this particular conscious system is not having an experience with less content; however, this fact in no way shows that there are no other conscious entities having other experiences, some with less content, e.g., the same scene in black and white.

The foregoing considerations show that the appeal to introspection amounts to an argument from ignorance, and the tendency to reject overlapping conscious entities appears unfounded.

But do we have independent support for overlap? Under some seemingly reasonable assumptions, it appears we have supporting evidence in medicine, science, and personal experience. Suppose we have good reason to believe some physical system is conscious and the question arises whether its right half is also conscious. While the overlap thesis can accept this possibility, IIT cannot. Now, suppose the left half of this system is temporarily disabled, and in that instant the active right half acts as if it just woke into consciousness, reporting no previous conscious experience. This would seem to count in favor of the view that the right half was not previously conscious, though it is by no means the only logical possibility.[17]

But consider another possible outcome. Suppose that, instead of acting as if it just woke into consciousness, the right half reports previous conscious experience and is able to note the changes it experiences due to having lost contact with the left half. It might, for example, demonstrate awareness of the loss of a faculty which had been controlled by the left half. Would this count in favor of the view that the right half had

[16] Split-brain cases provide instructive exceptions.
[17] Total amnesia induced by the disabling of the left half is one alternative.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

always been conscious? I suggest it would. Again, there are other logical possibilities, but imagine unconvinced IIT proponents explaining their case to the right half:

> Good morning, Right Half! We know you are conscious now, and we know you believe you have memories of your conscious past, but those memories are misleading. By the Exclusion Postulate, you were in fact not conscious before the left half was temporarily disabled. Moreover, by the Exclusion Postulate, when we now reactivate the left half, you will be conscious no more. Now, good night, Right Half… Good night.

Imagining the situation from a first-person perspective should sharpen the point. Imagine a moment of conscious experience that includes noticing that things are no longer as they have long been. Perhaps your capacity to speak is now gone, or you can no longer move one of your arms. If this experience is due to the fact that a portion of the brain has been disabled, then it is the surviving active portion of the brain which realizes this experience, but it, according to the Exclusion Postulate was not previously conscious. Nor will it be, should that portion of the brain be reactivated.

The direction this line of reasoning is going in should now be clear. For there are such conscious experiences in the real world. Consider, for instance, the experiences of stroke victims who, in the very midst of their stroke, suddenly become aware of the loss of a cognitive capacity, and note that this may involve the loss (perhaps only functional, and perhaps only temporary or even intermittent) of significantly large regions of the brain, regions that are proper parts of the largest brain region that was conscious prior to the stroke. Take neuroanatomist Jill Bolte Taylor's (2008) account of her stroke experience as she attempts to use her doctor's business card to call for help.

> To my astonishment, however, as I looked at the top card, I realized that although I retained a clear image in my mind of what I was looking for, I could not discriminate any of the information on the card in front of me. My brain could no longer distinguish writing as writing, or symbols as symbols, or even background as background. […]
>
> Dismayed, I realized that my ability to interact with the external world had deteriorated far more than I could ever have imagined.

Taylor's account is not simply of what it is like to be without important cognitive capacities such as the ability to recognize symbols which have been recognizable for the bulk of one's life; it is also an account of what it is like to *notice* that one has lost those capacities. And truly noticing the loss of a capacity would seemingly require a prior experience of having that capacity—however unattended that experience might have been.[18] Analogously, noticing that you've lost your wallet or that your arm has gone numb requires the prior experiences—however unattended—of having a wallet and of having an arm that isn't numb. We can ask, in Taylor's case of noticing the loss of the ability to recognize symbols, what exactly—by our best neuroscientific understanding—is the "noticer"?

Call the disabled region the *stroke region* and call the parts of the complement to the stroke region *survivor regions*. Which brain region is that region, the PSC in IIT terms, that can notice this kind of change, that can compare *what it is like now* with *what it was like then*? Presumably it is not any region which overlaps the stroke region, for the stroke region is not functioning. More plausibly, some currently active survivor region is not

[18] Consciously noticing change typically requires prior consciousness, though consciously noticing that one has just become conscious is an exception.

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

only that which experiences the loss of the stroke region but is also that which once enjoyed continuous seamless integration with that region. If so, then this survivor region existed as a conscious entity—a PSC—prior to the stroke. But it also overlaps a large incorporating region that existed as a conscious entity prior to the stroke, the larger incorporating PSC.[19] There would then be two overlapping conscious brain regions prior to the stroke, the region destined to survive and the larger, incorporating region destined to lose functioning parts due to stroke.

A stark case for overlap lies in the testimony of patients who have undergone the Wada test. The Wada test is a preoperative procedure which alternately anesthetizes brain hemispheres to locate language and memory capacities.[20] Some patients emerge from the Wada test with testimony of what it was like to lose one functioning hemisphere, then to regain it, then to lose the other, then to regain that other. Some of them report what it was like to lose their ability to name objects.

> My neurologist showed me a bunch of objects and photos prior to the test, then during the test he would show me something, ask me to identify it and whether he had showed it to me before. During the procedure they had me hold my arm straight up. For the right side of the brain I didn't notice anything different. For the left side - wow! When he showed me an object I looked at it and had that feeling you get when you can't think of a word, like it's on the tip of your tongue. Only that was true for all words - it was amazing! I had no words.[21]

The question the Wada test raises for our purposes is, "What part of the brain consciously experiences the loss (and return) of the left hemisphere, and what part consciously experiences the loss (and return) of the right?"[22] I suspect that the answer is, "Not the same part." Instead, the survivor region that witnesses the loss of the left hemisphere is not the survivor region that witnesses the loss of the right. After all, the right hemisphere (perhaps along with subcortical structures) can lose contact with the left hemisphere—discovering in that moment what it's like to lose all of one's words. But the right hemisphere (perhaps along with subcortical structures) cannot strictly lose contact with itself. And the wallet cannot lose the wallet.[23]

Surely many will resist this interpretation with the initially intuitive view that consciousness in the case of the Wada test "follows the survivor". Indeed, Tononi (2016) is clear that a conscious "complex" can move in the cerebral cortex.[24] Moreover, IIT would not be alone in observing that brain dynamics can be radically changed after the events under discussion. The objection to overlap here is that whatever surviving region is consciously experiencing the effects of loss of brain tissue or function, it has changed so significantly in this event that there is little reason to see it as having possibly been the same kind of thing, let alone an independently conscious system, prior to the loss. On this view, a "complex" (Tononi's term) does a kind of two-step as the anesthetic alternately disables the hemispheres. This complex contracts into one hemisphere,

[19] Another example might be found in split-brain patients, if we can recognize two PSCs differing by hemisphere but sharing subcortical structures.
[20] Snyder and Harris (1997), Meador *et al.* (1997), Cleveland Clinic (2021)
[21] Epilepsy.com/connect/forums/surgery-and-devices/wada-test-1
[22] Finding no difference, which is here reported for the right side, is also noticing something about a past and present experience: that they do not seemingly differ.
[23] One alternative hypothesis is that the conscious survivor is entirely in some subcortical region; the hemispheres are not at all conscious. While this would avoid the view that overlapping conscious systems exist in the hemispheres, it faces two challenges. First, the notion that the hemispheres are not at all conscious requires a radical revision of our current neuroscientific understanding of consciousness in the brain. Second, the possibility of overlap in this subcortical region remains. Another alternative to overlapping conscious systems is substance dualism.
[24] Godfrey-Smith (2020) provides a similar interpretation of the Wada test.

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

expands back to the whole brain, contracts into the other hemisphere, then expands back to the whole brain again. As mentioned, this "two-step" hypothesis (the "follow the survivor" hypothesis applied to the Wada test) is initially intuitive, but it faces two challenges.

First, if brain regions are the things which are conscious (the PSCs), then we appear to have brain regions which either falsely remember being conscious or accurately remember what it was like for another brain region to be conscious. On the first branch of this dilemma, the survivor region during a stroke only seems to remember once experiencing what it was like to, say, recognize numerals or produce coherent utterances in one's native tongue. But, as previously suggested, these memories are strictly false because, according to the two-step hypothesis, this region never had those conscious experiences. The second branch of this dilemma raises intriguing implications for Nagel's (1974) question of what it is like to be another conscious being. On this branch, a hemisphere would have some direct knowledge, in the form of very recent memory, of what it is like to be the whole cerebral cortex—exciting if true (for it points to a possible bridge we might try building toward answering Nagel's question) but not necessarily intuitive.[25] Thus, the two-step hypothesis entails either systematically false memories or a denial of the view that one conscious entity cannot know what it is like to be another conscious entity. Distinct regions, on this branch, can share qualia.

The second challenge involves parsimony. Those averse to a plurality of conscious systems might not find much solace the two-step hypothesis. Recall the view that whole-brain dynamics change so drastically that the brain region currently experiencing a stroke or other such event cannot possibly have been conscious prior to that event. In cases where experiences occur sometime after the fact, this is surely a reasonable concern, but conscious experience of loss can occur immediately after and even during these events, and these losses can be exceedingly fleeting. In the midst of a transient ischemic attack (TIA), whole-brain dynamics would need to alter in such a way that some functioning portion of the brain instantly becomes conscious, suddenly aware of its new symptoms such as weakness of limbs and trouble speaking. This newly minted conscious entity might not last long, for as a TIA progresses the borders between functioning and disabled regions can change. There are also subjects who have consciously experienced the temporary disabling of regions of the cortex by transcranial magnetic stimulation (TMS), reporting in the moment, for example, what it is like to have temporary blindsight.[26] The disabling TMS pulse and the experience can last milliseconds. The "follow the survivor" alternative to overlap implies that in such cases numerous conscious systems spring into and out of existence, some experiencing conscious life for a mere fraction of a second.

The initially intuitive view that consciousness can follow the surviving brain regions evidently has counterintuitive implications. The view that conscious systems can overlap avoids them. Furthermore, under the reasonable assumption that at least some surviving portions of a brain would have accurate memories of some of their past conscious experiences had they had any, we appear to have a testable hypothesis, one which has been tested and substantially confirmed. For instance, at least some patients of the Wada test should indicate realizing that they have lost faculties they remember having, as some do, instead of experiencing something akin to waking with no memories at all. And some of those who have had a TIA

[25] This option exists for the overlap view as well. Reasonable possibilities for overlap are that these brain regions accurately remember being conscious or accurately remember what it was like for another brain region to be conscious.
[26] See Koenig and Ro (2019). Patients with blindsight (whether natural or artificially induced) report having no conscious visual experience in some or all of their visual receptive field but can nevertheless process visual stimuli regarding color, location, shape, path of motion, and orientation in the blind field.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

should report noticing the change, as some do, instead of experiencing something akin to waking with no memories at all.

Nevertheless, the overlap view remains counterintuitive. Not only does it violate our common aversion to the overlap *per se* of conscious systems, but the overlap view's further commitments to plurality and composition, as previously described, are themselves counterintuitive. Finally, the sheer number of such systems is astronomical. After all, because there are very many brain regions whose loss can be survived by another region, the argument for overlap generalizes in a way that will surely sound absurd at first to many researchers, even those who are willing to accept one independently conscious system for each hemisphere. It would be coy to ignore this generalization, and its personal implications.

> Right now, as "you" (a pronoun whose conventions are especially challenged at this point) are reading this passage, "parts of you" are experiencing what it is like to be conjoined and interacting with other parts which together compose the entity that is reading the passage. These are the parts which would survive as conscious noticers of the loss of other parts, were those other parts to cease functioning or to become disconnected.

We can admit that the existence of conscious overlapping systems in the human brain is strange, counterintuitive, or even bizarre while simultaneously favoring or even embracing it, if it is among the least bizarre theories.[27] Here we seem to have a choice between a view according to which conscious entities overlap and a view according to which consciousness is not intrinsic, sometimes coming and going (possibly with a mass of false memories) because of mere comparisons.

Finally, it is also worth considering whether overlap of this kind is not at all bizarre or even counterintuitive once we consider the broader scientific context, for our best physical theories explicitly admit of properties and quantities which hold of overlapping things. A physical system can have temperature, mass, volume, energy, and charge even when certain of its proper parts also have temperatures, masses, volumes, energies, and charges.[28] When scientists measure and investigate such properties, there is no assumption that the "winner" must be determined. Why should it be so different with consciousness?

5. Conclusion

We have seen that if consciousness is intrinsic in the sense used here, then consciousness is not Φ^{max}. To preserve intrinsicality, and to avoid zombies and the other surveyed consequences, IIT must reject its Exclusion Postulate. This conclusion is by no means devastating to the promising and plausible view that consciousness is somehow a matter of a special kind of causal integration, but it does suggest that IIT should abandon its search for the "winner" and admit that conscious physical entities can overlap. This is not all.

Recall that, as previously mentioned, rejecting the Exclusion Postulate is insufficient for preserving intrinsicality. Tononi accepts that Φ itself is also extrinsic because the amount of Φ in a system depends on

[27] See Schwitzgebel (2014).

[28] The fact that some of these properties are extensive (changing as the amount of matter changes) and some are intensive (remaining unchanged as the amount of matter changes) may cause confusion. Volume is an extensive property, while temperature is an intensive property, but a system can have properties of both kinds while various of its parts also have those properties.

Journal of Consciousness Studies 28 (11-12): 31-53. 2021.

background conditions which could be different, thereby determining a different value of Φ. Integration itself, then, is not intrinsic, so rejecting the Exclusion Postulate is insufficient to preserve intrinsicality.

Critics may shrug: So much the worse for IIT. However, IIT proponents seeking to preserve the intrinsicality of consciousness might turn their attention to the intrinsic property of a system which determines that system's Φ relative to background conditions. After all, if Φ measures something promising about consciousness, but Φ is extrinsic, then, whether consciousness is intrinsic or not, it would be fruitful to ask what invariant property of a system determines that system's various degrees of Φ relative to various background conditions. Φ in this sense is like weight. An object's weight varies according to the gravitational field the object is in, making weight an extrinsic property. As we discovered, mass is the property that determines an object's weight relative to a gravitational field. And upon finding that mass varies according to frame of reference, we took that step again, finding nonrelativistic mass to be that invariant which determines relativistic mass. Analogously, Φ's extrinsic nature should point IIT researchers in the direction of discovering its intrinsic base: that invariant which determines the Φ of a system relative to background conditions.

Whether such a property is plausibly consciousness is another matter, but unless consciousness is extrinsic, this base property would at least be a candidate, unlike Φ or Φ^{max}. Moreover, even if this base property is not identical to consciousness, it nevertheless plays a significant role in determining and explaining the causal integrity and dispositions of conscious physical systems. Identifying this invariant would be a step forward.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

Acknowledgements

I am grateful to Francis Fallon, John A. Keller, Jo Edwards, Eric Schwitzgebel, Carlos Montemayor, Judy Lee, and two anonymous reviewers for insights and work that substantially improved this manuscript. Thanks also to members of the Redwood Center for Theoretical Neuroscience at UC Berkeley, the Science and Consciousness Conference in Tucson, Arizona, and the Initiative for a Synthesis in Studies in Awareness at the Okinawa Institute of Science and Technology Graduate University, Japan.

References

Aaronson, S. (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.

Block, N. (1978) Troubles with Functionalism, *Minnesota Studies in the Philosophy of Science*, 9, pp. 261-325.

Casati, R. and Varzi, A. C. (1999) *Parts and Places: The Structures of Spatial Representation*, Cambridge (MA): MIT Press.

Cleveland Clinic (2021) *Wada Test*, https://my.clevelandclinic.org/health/diagnostics/17628-wada-test [1 May 2021]

Epilepsy.com (2021) *The Wada Test*, https://www.epilepsy.com/connect/forums/surgery-and-devices/wada-test-1 [1 May 2021]

Fallon, F. (2020) Integrated Information Theory, Searle, and the Arbitrariness Question, *Review of Philosophy and Psychology* 11 (3), pp. 629-645.

Godfrey-Smith, P. (2020) *Metazoa*, New York: Farrar, Straus and Giroux

Horgan, J. (2015) Can Integrated Information Theory Explain Consciousness? *Scientific American*, https://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/ [4 Jan 2021]

Kim, J. (1987) 'Strong' and 'Global' Supervenience Revisited, reprinted in Kim 1993, pp. 79–91.

Kim, J. (ed.) (1993) *Supervenience and Mind: Selected Philosophical Essays*, Cambridge: Cambridge University Press.

Koch, C. and Tononi, G. (2013) Can a Photodiode Be Conscious? *New York Review of Books*.

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

Koenig, L. and Ro, T. (2019) Dissociations of conscious and unconscious perception in TMS-induced blindsight, *Neuropsychologia*, 128, pp. 215-222.

Lewis, D. (1983) Extrinsic Properties, *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, Vol. 44, No. 2, pp. 197-200.

***Journal of Consciousness Studies* 28 (11-12): 31-53. 2021.**

Meador, K. J, Loring, D. W., Lee, G. P., Nichols, M. E., Moore, E. E., Figueroa, R. E. (1997) Level of Consciousness and Memory during the Intracarotid Sodium Amobarbital Procedure, *Brain and Cognition*, 33, pp. 178–188.

Merricks, T. (2003) Maximality and Consciousness, *Philosophy and Phenomenological Research*, Vol. 66, No. 1, pp. 150-158.

Mørch, H. H. (2019) Is Consciousness Intrinsic? A Problem for the Integrated Information Theory, *Journal of Consciousness Studies*, Vol. 26, Numbers 1-2, pp. 133-162.

Nagel, T. (1974) What is it like to be a bat? *Philosophical Review*, LXXXIII (4), pp. 435–450.

Oizumi, M., Albantakis, L., Tononi, G. (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0, *PLOS Computational Biology*, https://doi.org/10.1371/journal.pcbi.1003588

Putnam, Hilary (1975) *Mind, Language and Reality: Philosophical Papers* Cambridge University Press, 434.

Schwitzgebel (2014) The Crazyist Metaphysics of Mind, *Australasian Journal of Philosophy*, 92 (2014), 665-682.

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

Schwitzgebel (2020) The Nesting Problem for Theories of Consciousness, *The Splintered Mind*, http://schwitzsplinters.blogspot.com/2020/11/the-nesting-problem-for-theories-of.html, [1 May 2020]

Searle, J. (2013) Can information theory explain consciousness? *New York Review of Books*.

Sider, T. (2003) Maximality and Microphysical Supervenience, *Philosophy and Phenomenological Research*, Vol. 66, No. 1, pp. 139-149.

Snyder, P. J. and Harris, L. J. (1997) The Intracarotid Amobarbital Procedure: An Historical Perspective, *Brain and Cognition*, Vol. 33, pp. 18–32.

Taylor, J. B. (2008) *My Stroke of Insight*, New York: Viking

Tononi, G. and Koch, C. (2015) Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society* B.

Tononi, Giulio. (2004) An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5:42.

Tononi, G. (2015) Integrated information theory, *Scholarpedia*, http://dx.doi.org/10.4249/ scholarpedia.4164 [11 Dec 2020].

Tononi, G., Boly, M., Massimini, M., Koch, C. (2016) Integrated information theory: from consciousness to its physical substrate, *Nature Reviews: Neuroscience*, Vol. 17, MacMillan Publishers, pp. 450-461.

Unger, P. (1980) The Problem of the Many, *Midwest Studies in Philosophy*, 5, pp. 411–67.

Unger, P. (2004) The Mental Problems of the Many, *Oxford Studies in Metaphysics*, Vol. 1.

Van Stekelenburg, T. and Edwards, J. (2020) Why Integrated Information Theory Must Fail on its Own Causal Terms, *Journal of Consciousness Studies*, **27**, No. 7-8, pp. 144-164.