PhiMiSci
Philosophy and the Mind Sciences

On the dangers of conflating strong and weak versions of a theory of consciousness

Matthias Michel[a] (matthias.michel.curtil@gmail.com)
Hakwan Lau[b] (hakwan@gmail.com)

Abstract
Some proponents of the Integrated Information Theory (IIT) of consciousness profess strong views on the Neural Correlates of Consciousness (NCC), namely that large swathes of the neocortex, cerebellum, basal ganglia, thalamus, olfactory bulb, and the so-called limbic system, are all not essential for any form of conscious experiences. We argue that this claim derives from a strong version of the theory, according to which the NCC is supposed to constitute conscious experiences. On a weaker version of the theory, IIT only provides what we call a marker of consciousness. We surmise that a conflation between strong and weak versions of the theory has led these researchers to adopt definitions of NCC that are inconsistent with their own previous definitions, inadvertently betraying the promises of an otherwise fruitful empirical endeavour.

Keywords
Consciousness · Integrated information theory · Neural correlates of consciousness

This article is part of a special issue on "The Neural Correlates of Consciousness", edited by Sascha Benjamin Fink.

1 NCCs, markers, and constituents

> *We are not receptive to physicists trying to apply exotic physics to the brain, about which they seem to know very little, and even less about consciousness. – (Crick & Koch, 2003)*

Identifying the neural correlates of consciousness (NCC) has been a central research program in consciousness science for decades. In common neurobiological

[a]Centre for Philosophy of Natural and Social Science, London School of Economics and Political Science
[b]Department of Psychology & Brain Research Institute, UCLA; Department of Psychology & State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54
 https://philosophymindscience.org ISSN: 2699-0369

language, the key term would have been 'mechanisms' rather than 'correlates.'[1] The term 'correlates' was carefully chosen to remain neutral on conceptual issues regarding the exact metaphysical relation between consciousness and the NCCs (Crick & Koch, 1990). It was recognized early on that such questions are difficult, and best set aside until we have clearer answers on the NCC.

In this article, we introduce a conceptual distinction between NCCs, *markers* of consciousness, and *constituents* of consciousness.[2] We argue that a failure to distinguish between these different concepts is limiting the progress of the NCC project. We will illustrate this point with the case of the Integrated Information Theory (IIT) of consciousness.

An NCC is "the minimal set of neuronal events and mechanisms jointly sufficient for a specific conscious percept" (Chalmers, 2000; Koch, 2004, p. 16). That is, NCCs are the *minimal* neural difference makers that are jointly *sufficient* for a mental state to be conscious, rather than unconscious. This is just another way of saying that, all other things being equal, if a mental state is unconscious, activation of the NCC should be sufficient for making it conscious.

We distinguish the NCC from the *constituents* of consciousness. If they exist, constituents of consciousness are neural, or physical states that are *identical* with consciousness. This means that the constituents of consciousness should be sufficient – *and necessary* – for consciousness, just as H_2O is the constituent of water.

Research has focused on finding NCCs rather than constituents of consciousness because the *necessity* condition is generally considered too strong (Chalmers, 2000). Indeed, if a neural state is not only sufficient, but also absolutely *necessary* for a specific conscious experience to occur, it means that that experience is not multiply realizable (Chalmers, 2000; Michel et al., 2018; Morales & Lau, 2018). That is, as in the case of water being identical to H_2O, we cannot replace hydrogen or oxygen with something else to produce water. H_2O is the only possible recipe. This claim would, in effect, be contradictory with the widely acknowledged phenomenon of *degeneracy* in biology (Edelman & Gally, 2001; Tononi et al., 1999) – which states that given a context, the same biological function can be carried out by different substrates, giving the impression of redundancy. If the same conscious experience could be realized with slightly different neuronal ensembles which may

[1] Following Illari & Williamson's 'consensus concept' of mechanism, a mechanism for a phenomenon is defined here as a set of "entities and activities organized in such a way that they are responsible for the phenomenon" (Illari & Williamson, 2012, p. 120). Identifying consciousness mechanisms requires one to understand what the mechanisms *do* for consciousness exactly (i.e. their functions). Identifying markers, correlates, or even constituents, does not require one to do so. We will not focus on mechanisms in this article (for more on mechanistic explanations in consciousness science, see Miracchi (2017)).

[2] A constituent of consciousness is different from what is sometimes called a 'core NCC'. A core NCC is "the part of the total NCC that distinguishes one conscious state from another – the rest of the total NCC being considered as the enabling conditions for that conscious experience" (Block, 2005, p. 47). A core NCC is not necessary and sufficient for a given conscious experience – it is only sufficient when combined with some enabling conditions. By contrast, a constituent of a conscious experience is necessary and sufficient for that conscious experience to occur.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

largely overlap but are not *exactly* identical, an NCC at this level cannot be considered strictly necessary for a particular experience. We will come back to this point below.

Lastly, we distinguish constituents and NCCs from the *markers* of consciousness. By markers we mean general evidence that can be used to determine whether subjects have conscious mental states or not. In that sense, NCCs can be used as markers of consciousness, because if a person is able to entertain a specific conscious experience, it demonstrates that the person is conscious in at least some limited sense. But not all markers of consciousness are NCCs, or constituents of consciousness. These markers may be very general and might not specifically reflect the mechanisms directly responsible for consciousness. Some markers of consciousness might correspond to 'pre-requisites' or 'consequences' of consciousness (Aru et al., 2012).

To illustrate this distinction, let us consider the fact that when one is conscious rather than in a coma, one is more capable of producing behavior, thoughts, and memory. Therefore, a neural marker for sophisticated cognition may generally be used as a marker for consciousness. But such a marker may not be the *minimally* sufficient condition for specific conscious experiences; one may need much less for a specific single experience to arise. And such a marker may also not be necessary for conscious experiences; one might be able to entertain a specific experience without sophisticated cognition. Still, it could be that, *in general*, one can pragmatically use a neural marker for sophisticated cognition as a marker for consciousness, as consciousness and cognition might generally correlate, at least in humans.

Clearly distinguishing between markers, NCCs, and constituents of consciousness is important for all theories of consciousness: conflating these three notions might create confounds. For instance, the P3b wave was considered an NCC by proponents of the global workspace theory (Dehaene & Changeux, 2011). Recent experiments, however, have shown that the P3b is neither necessary (Cohen et al., 2020; Pitts et al., 2014) nor sufficient (Silverstein et al., 2015) for consciousness. While one can interpret the P3b as a reliable *marker* of consciousness in some cases, it cannot be interpreted as an NCC, or a constituent of consciousness. Clearly distinguishing between markers and NCCs in this case shows that an electrophysiological signal, while not being mechanistically relevant for consciousness, can remain practically useful (e.g. in clinical settings) as a marker of consciousness (Faugeras et al., 2012).

While we maintain that the distinction is important for all theorizing about consciousness, what follows is a case study of how the distinction plays out for a specific theory: the integrated information theory of consciousness (IIT). We argue that a failure to carefully distinguish between NCCs, markers, and constituents of consciousness, is at the heart of a conflation between a strong version of IIT, and a weaker, empirical version.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

2 Empirical vs Fundamental IIT

The recent 'rise' of IIT is somewhat paradoxical. On the one hand, the theory has been promoted with an unusual level of enthusiasm. For example, it has been claimed that IIT is "currently *accepted* as one of the most compelling explanations about what consciousness is"[3] (italics ours), that it is a "gigantic step in the final resolution of the ancient mind-body problem" (Koch, 2004), and even that it yields "a new kind of scientific spirituality."[4]

On the other hand, arguments and criticisms against the theory abound (Barrett & Mediano, 2019; Bayne, 2018; Cerullo, 2015; Doerig et al., 2019; Pautz, 2019). To answer one of those criticisms, proponents of IIT ended up acknowledging that, according to the theory, even a set of 'inactive' logic gates would be conscious (Aaronson, 2014b, 2014a; Tononi, 2014). Such possibilities have been openly denounced as "untestable" by many (Michel et al., 2019)[5]. Overall, the theory has not impressed active researchers as much as non-experts outside the field (Michel et al., 2018).

We suspect that this striking difference in opinions is due to a conflation between two versions of the theory, which, following Mediano et al. (2019), we call *Empirical* IIT and *Fundamental* IIT. While the former may have some merits, we are unsure about the latter.

Empirical IIT is the view that measures of integrated information in brain networks – a specific subtype of complexity measure – can be used to detect states of consciousness, i.e., whether subjects have subjective experiences (such as during wakefulness, or dreams), or not (such as being in a coma, anesthetized, or dreamless-sleep). That is, Empirical IIT takes integrated information to be a *marker* of consciousness.

According to this view, the choice of an exact quantitative measure of integrated information is an empirical matter. For instance, if it turns out that some measure of integrated information does a better job at categorizing states of consciousness accurately, this would count as empirical data in favor of using that measure. Overall, despite some contradictory evidence (Noel et al., 2019; Sasai et al., 2016; Tajima et al., 2015),[6] there is support for Empirical IIT (Barttfeld et al., 2015; Bodart et al., 2017; Casali et al., 2013; Casarotto et al., 2016; Demertzi et al., 2019; Ferrarelli et al., 2010; Rosanova et al., 2012; Ruiz de Miras et al., 2019; Sarasso et al., 2015; Tagliazucchi et al., 2013). We can consider Empirical IIT scientifically plausible.

[3] See: https://qz.com/709969/2300-years-later-platos-theory-of-consciousness-is-being-backed-up-by-neuroscience/

[4] https://www.huffpost.com/entry/post_b_8160914

[5] This point was also made in a recent letter to the NIH, by dozens of researchers in the field, available online here: https://tinyurl.com/y5wokv9g.

[6] See http://inconsciousnesswetrust.blogspot.com/2017/08/how-to-make-iit-and-other-theories-of.html for discussion as to why Sasai et al. (2016) and Tajima et al. (2015), and some other studies are considered potentially contradictory evidence against Empirical IIT.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

On the other hand, Fundamental IIT is the view that a specific form of complexity (integrated information) is *identical* with consciousness (Oizumi et al., 2014). That is, Fundamental IIT posits that integrated information is the *constituent* of consciousness. Specific conscious experiences are also exactly identified with specific states of a network with particular patterns of integrated information (Tsuchiya et al., 2015; Tsuchiya, 2017).

As such, Fundamental IIT does not simply posit that integrated information is necessary or sufficient for consciousness, or that integrated information is a marker of consciousness. Instead, consciousness is *identified* with a system having the features described by the 'postulates' of the theory. This identity claim is explicit in the writings of proponents of IIT:

> according to IIT, there is an identity between phenomenological properties of experience and informational/causal properties of physical systems (see [11] and [19] for the importance of identities for the mind-body problem). The central identity is the following: The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience. (Oizumi et al., 2014, p. 3)

> Importantly, IIT claims that a *quale* in the broad-sense is *identical* to a MICS, generated by a particular subset of the neural system, for example, a thalamo-cortical system excluding cerebellum. In other words, IIT proposes a mapping between a certain mathematical structure, which is derived from connectivity and a state of a certain subset of the neurons in the brain, and the particular quale that a subject of the brain is experiencing. (Tsuchiya et al., 2015, p. 3)

As Fallon (2015) notes, proponents of IIT also routinely offer "analogies to other fundamental physical properties. Consciousness is fundamental to integrated information in the same way as it is fundamental to mass that space-time bends around it". In summary, proponents of IIT subscribe to an identity between types of experiences (shapes in 'qualia-space') and types of patterns of integrated information (shapes in a 'cause-effect space'), realized in physical causal structures (Tsuchiya, 2017). It follows that identifying the causal structure of a physical system, and deriving its integrated information, is sufficient to determine which type of experiences that system has, and vice versa.

Importantly, causal structures can be realized in various physical substrates. Physical systems that are composed of different stuff can have the same causal structures and thus realize the same patterns of integrated information. As Tsuchiya (2017) notes:

> essential relationships in IIT are those between consciousness and mathematical structures derived from the physical substrates, not between consciousness and matter as is usually debated in philosophy. This means that two distinct physical substrates can generate identical consciousness.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

Some could argue that this aspect of IIT makes the theory more similar to functionalism than identity theory, given that functionalism has historically been associated with multiple realizability (Polger & Shapiro, 2016). However, IIT is very far from functionalism, for at least two reasons. First, there is a crucial sense in which IIT does *not* accept multiple realizability: a given type of experience cannot be realized by different types of physical causal structures, even if those causal structures themselves can be realized in different substrates. For instance, the *causal structure* realizing the experience of seeing something red can be realized in a brain or in swiss cheese, and the substrate will have the experience of seeing something red as long as the right *causal structure* is maintained. But the same experience cannot be realized by *different causal structures*, even with the *same substrate*. Two brains cannot realize the same experience by instantiating different causal structures. In that last sense, conscious experiences are not multiply realizable, according to IIT.

Second, and more importantly, conscious experiences are not tied in any way to the realization of psychological *functions*. This is made clear by the fact that feedforward networks could realize the exact same functions as the human brain without being conscious, according to IIT (Doerig et al., 2019). What matters for consciousness is not *what* functions are realized – since even a set of inactive logic gates can have conscious experiences (Aaronson, 2014b), but only *how* functions are physically realized. For this reason, proponents of Fundamental IIT are anti-functionalists, as Tononi & Koch (2015) write: "in sharp contrast to widespread functionalist beliefs, IIT implies that digital computers, even if their behaviour were to be functionally equivalent to ours, and even if they were to run faithful simulations of the human brain, would experience next to nothing.".

As such, in contrast with Empirical IIT, which considers integrated information as a *marker* of consciousness, Fundamental IIT postulates an *identity* between patterns of integrated information and conscious experiences. In addition, according to Fundamental IIT, the exact measure of integrated information to adopt is not a matter to be determined by the data. Rather, it stems from some assumptions, or axioms, that are meant to be self-evidently true (Oizumi et al., 2014; Tononi et al., 2016; Tononi & Koch, 2015). From there, one is supposed to mathematically derive the measure of integrated information that is relevant for consciousness. We will argue that Fundamental IIT is not in line with current scientific knowledge and practice. It may also jeopardize the NCC project.

3 Conflating general markers and constituents

Let's assume that some versions of the integrated information measure can be successfully used as markers of states of consciousness, as claimed by Empirical IIT. Importantly, we have to recognize that this would be entirely compatible with the rejection of Fundamental IIT.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

That is to say, data in favor of Empirical IIT may be compatible with other theories, such as the Global Workspace Theory (GWT)[7]. According to GWT, consciousness results from the broadcast of information to a wide variety of neuro-cognitive modules through a "global workspace". If one supposes that this global broadcast is best achieved in a network of high complexity, GWT can account for currently available evidence in support of Empirical IIT, without invoking the controversial conjectures of Fundamental IIT.

Also, that *some* measures of integrated information can be used as *markers* of states of consciousness does not directly support the claim that *integrated information* is *identical* with consciousness. Because the exact calculation of the degree of integrated information in real biological systems is computationally challenging (Barrett & Mediano, 2019),[8] studies currently rely on proxy measures. Due to the approximate nature of the measurements, the relevant results cannot support claims concerning constituents, rather than markers. That is, they support Empirical, not Fundamental IIT.

On this note, we remark there are many different approximated measures of integrated information, and they do not necessarily empirically converge (Mediano et al., 2019). Given the same data Fundamental IIT can both be supported or falsified by these measures, depending on which approximation we choose to adopt.[9] Relatedly, there is currently a debate as to whether Fundamental IIT is empirically falsifiable at all (Doerig et al., 2019).

Finally, and perhaps most importantly, in order to interpret the successful results of integrated information measures as support for the claim that consciousness is *identical* with integrated information, proponents of Fundamental IIT would have to rule out important confounds. For instance, a wide variety of *cognitive* capacities also differ between wakeful and unconscious subjects. As such, it could be that integrated information measures reflect *these* differences, instead of differences in consciousness *per se*. As a result, to use data obtained with integrated information measures of wakeful versus unconscious subjects to support Fundamental IIT, proponents of Fundamental IIT would have to provide good reasons for believing that integrated information indexes *consciousness* and not *cognition*, or any other factor that varies between wakeful and unconscious individuals.

To be clear, again, this latter argument applies to Fundamental IIT, but not to Empirical IIT. As long as some integrated information measures *do* correlate with

[7] This is apparently acknowledged by some proponents of IIT (Koch says that "The global workspace theory and integrated information theories are not mutually exclusive", see: https://www.livescience.com/47096-theories-seek-to-explain-consciousness.html)

[8] As remarked by Barrett & Mediano (2019), this is because "the computation time required to compute [integrated information] grows faster than exponentially with the number of system components" (p.1).

[9] See https://jakerhanson.weebly.com/blog/my-graduate-experience-with-integrated-information-theory-iit for a useful discussion on the difficulty in pinning down a complexity measure for this purpose.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

differences in states of consciousness, whatever the reason behind this correlation, those measures can be used as *markers* of consciousness. It is an entirely different matter, however, to claim that integrated information should be *identified* with consciousness, as do proponents of Fundamental IIT.

Therefore, if we do not confuse markers and constituents of consciousness, it should be clear that reasons for finding Empirical IIT appealing generally do not extend to Fundamental IIT.

4 Conflating constituents and NCCs

To make claims about the constituents of consciousness, we need to do more than just ruling out some confounds. Identity relationships like that between water and H_2O are very strong statements to make. To do so, we first need to have a very good understanding of the relevant substrates (hydrogen and oxygen in this case), their causal properties, how they interact, as well as to make sure that they are necessary. However, in the case of Fundamental IIT, it is unclear what kind of physical states are supposed to be identical to consciousness. If the physical substrate itself is not clearly identified, IIT is very far from establishing a convincing identity statement.

Indeed, *what substrate* is supposed to be identical with consciousness, according to Fundamental IIT? According to the theory, the measure of integrated information to be assessed concerns some 'nodes' of a physical system that are connected with each other. These nodes are either on, or off. They are either connected to other nodes or not. Of course, we already know that neurons are much more complex than simple on/off nodes, as they show degrees of intensity of firing (rate), different dynamics in firing, different types of connections with each other (Kandel et al., 2012). As such, proponents of Fundamental IIT acknowledge that these 'nodes' likely operate at a "different level."[10] That is, they are most probably *not* to be identified with necessarily neurons, as one may intuitively think. As a result, it is unclear *what* the substrate of consciousness is supposed to be, according to Fundamental IIT. This has the downside that it is not clear where researchers should look when trying to validate or invalidate the claims of Fundamental IIT.

But let us assume that some – possibly sub-neuronal – substrate identified by Fundamental IIT does correlate with consciousness. Even so, such correlation will

[10]There is a sense in which the theory captures more than simple on/off states in the nodes, because formally the theory applies to general stochastic systems that are parameterised by probability distributions and transition matrices. But ultimately, even with stochasticity, the possible outcomes are still on versus off. This does not capture the complex known dynamics of neurons. Accordingly, the intuitive examples given to explain the theory also tend to concern simple binary on/off nodes. The fact that these nodes are not really intended to plausibly model neurons has been confirmed in personal communications with the proponents of the theory. Personal communications with Andrew Haun, and Masafumi Oizumi. See a conversation between Hakwan Lau, Masafumi Oizumi, and Richard Brown, available here: https://www.youtube.com/watch?v=8AP6YQrwaN0. Discussion of this point is roughly between 21m38s – 31m28s.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

only support that these constituents may be candidates for the NCC, but not that they are identical to conscious experiences. To make the latter claim, one would need to show that they are absolutely necessary, just like we cannot replace oxygen or hydrogen with something else if we want to produce water.

This kind of strong identity relationships are unlikely to be found in most areas of biology, and neuroscience in particular, in which we generally try to identify *mechanisms* (Craver, 2007). Indeed, if a given neural state N is *necessary* for a given conscious experience C, C cannot obtain unless N does. This implies that C could not be realized by a very slightly different neural state, N_2. Let's say that N involves tens of thousands of 'nodes'. N_2 would be considered different in this context even if only the connection between two out of these nodes were modified, with everything else being identical to N.[11] This would be broadly inconsistent with the phenomenon of biological degeneracy, which is particularly prominent in highly complex systems – a point with which proponents of IIT should be familiar (Edelman & Gally, 2001; Tononi et al., 1999).

We therefore advocate maintaining the traditional definition of NCC. According to this definition, an NCC is a minimal set of neural activity sufficient for having a conscious experience (Chalmers, 2000). It differs from the 'non-traditional' notion used by proponents of IIT – as we show in the next section, which implies identity between physical causal structures and conscious experiences. If we are to make such a radical modification of the definition, to move from mere correlation to absolute identity, we should do so explicitly. Unfortunately, we suspect this change in definition has recently slipped into current debates tacitly.

5 NCC confused

One of the most striking claims made by proponents of IIT, beside panpsychism (Tononi, 2014), may be their views on the NCC. In particular, Christof Koch and colleagues have made the strong claim that most areas outside of a putative region in the *'posterior cortex'* are not home to the NCC. The excluded regions presumably include the insula, amygdala, olfactory bulb, basal ganglia, thalamus, and different parts of the prefrontal cortex (PFC). In particular, this is meant to concern almost *all* conscious experiences, not just conscious perception (Koch et al., 2016; Tononi et al., 2016).

[11] This is not to say for any N and N_2, the two must always lead to the same outcome on consciousness. The point about degeneracy is not that it happens 100% of the time. Rather, that it can sometimes happen at all means that the identity claim involving necessity cannot be logically correct. Chalmers (2000) discussed related cases, as well as other types of 'redundancy'. Overall, there may be more complexity to these issues here. For example, when we usually have two neural states, each independently sufficient for a subjective experience, if one state is abolished, one can consider the 'background conditions' changed, in which case, no strong definite predictions could be made. But in any case, it is clear that any reasonable and consistent interpretation of the original definition of NCC is inconsistent with the quote given in footnote 12.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

We remark that this claim may seem grossly incompatible with standard textbook knowledge, especially regarding conscious experiences of hunger, emotions, pain, intentions, thoughts, etc. But perhaps analyzing the case of the role of PFC for conscious visual perception is more illuminating still, because Koch himself used to hold exactly the opposite view. Here, we suggest that his more recent view on this matter is likely the result of a new commitment to the search for the *constituents* of consciousness, instead of the search for the NCCs.

Curiously, Koch's new justification for excluding PFC from the NCC is largely based on the very same data that previously led him to accept that the PFC was an NCC for conscious visual experiences. Indeed, Crick & Koch (Crick & Koch, 1998, p. 103) hypothesized that patients with bilateral lesions to the PFC may be able to respond unconsciously to visual stimuli, without being conscious of those stimuli. However, some well-known cases of bilateral lesions to the PFC had long been taken as indicating that patients with these lesions were not blind (Brickner, 1952). At that time, Koch concluded that these cases were not decisive for settling the debate on whether PFC was a neural correlate of consciousness (Crick & Koch, 1998, p. 103; Koch, 2004).

To our minds, these cases remain as indecisive as they were two decades ago, in part because the completeness of these lesions remains a matter of dispute (Odegaard et al., 2017). Meanwhile, it is now known that even in unilateral PFC lesions, there are in fact specific perceptual deficits (Fleming et al., 2014).

But more importantly, in recent correspondence, Koch insists that the lack of complete abolishment of conscious perception in these unilateral cases is evidence for writing off the entire PFC. Specifically, according to his current thinking, disruption of any part of the NCC should "necessarily" cause changes in conscious experience.[12]

This 'new' definition of the NCC, involving necessity, is in direct contradiction with the original definition, as we explained in the opening section. For example, a group of neurons in the right PFC may be on their own *minimally sufficient* for a conscious experience to occur. But once lesioned, neurons in the left PFC may take over to perform the same function. This kind of dynamic reorganization of function in the PFC has also been empirically demonstrated (Voytek et al., 2010).

However, if one is committed to Fundamental IIT, specifically the claim of neural *identity*, one may well hold that degeneracy does not apply to consciousness. This seems to be the most charitable interpretation of the inconsistency between Koch's previous and current views on the NCC.

So, conflating the constituents and the NCCs probably has already had an impact on how current research and debates are conducted.

[12] In an email correspondence dated June 24, 2018, Koch wrote: "*Any* change in this NCC (via a different stimulus, or causal intervention such as TMS, optogenetics, drugs etc) will, *of necessity*, change the character of the experience (including having no experience). If the background conditions change but the NCC does not, the experience will likewise not change." (italics ours). Quoted with explicit permission.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

6 Concluding remarks

We have compared two versions of a theory of consciousness. While there is some empirical support for the weaker version (Empirical IIT), going from there to a much stronger version (Fundamental IIT) seems to require an unscientific leap of faith. The points we have made here are not specific to IIT *per se*. In general we should not make weighty claims beyond what is warranted by evidence. In analyzing the case of IIT, we also illustrate two culprits limiting progress in consciousness science.

The first mistake is to interpret evidence in favor of using a neural state as a *marker* of consciousness as evidence for the *mechanistic relevance* of this neural state for consciousness. This point applies specifically to the study of states of consciousness, in which we might often be too quick in identifying markers of consciousness with NCCs.

The second mistake is to confound NCCs with neural *constituents* of consciousness. As we have seen, doing so could lead to the rejection of the well-known biological phenomenon of degeneracy, and to controversial interpretations of lesion studies. We would do well not to let our theoretical commitments derail the NCC project.

Acknowledgments
We thank two anonymous reviewers for their comments. We thank David Chalmers, Christof Koch, and Masafumi Oizumi for helpful discussions.

References

Aaronson, S. (2014a). *Giulio Tononi and Me: A Phi-nal Exchange*. https://www.scottaaronson.com/blog/?p=1823

Aaronson, S. (2014b). *Why I am not an integrated information theorist (or, the unconscious expander)*. www.scottaaronson.com/blog/?p=1799

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience and Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Barrett, A. B., & Mediano, P. A. M. (2019). The phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2).

Barttfeld, P., Uhrig, L., Sitt, J. D., Sigman, M., & Jarraya, B. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences*, *112*(37), E5219–E5220. https://doi.org/10.1073/pnas.1515029112

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1), 1–8. https://doi.org/10.1093/nc/niy007

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Sciences*, *9*(2), 46–52. https://doi.org/10.1016/j.tics.2004.12.006

Bodart, O., Amico, E., Wannez, S., Gomez, F., Casarotto, S., Rosanova, M., Casali, A. G., Gosseries, O., Laureys, S., Massimini, M., & Martens, G. (2017). Global structural and effective connectivity in patients with chronic disorders of consciousness. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, *10*(2), 353. https://doi.org/10.1016/j.brs.2017.01.035

Brickner, R. M. (1952). Brain of patient A. After bilateral frontal lobectomy; Status of Frontal-Lobe Problem. *Archives of Neurology and Psychiatry*, *68*(3), 293–313. https://doi.org/10.1001/archneurpsyc.1952.02320210003001

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and

behavior. *Science Translational Medicine*, *5*(198), 198ra105 LP–198ra105. http://stm.sciencemag.org/content/5/198/198ra105.abstract

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., G. Casali, A., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*, *80*(5), 718–729. https://doi.org/10.1002/ana.24779

Cerullo, M. A. (2015). The problem with phi: A critique of integrated information theory. *PLoS Computational Biology*, *11*(9), 1–12. https://doi.org/10.1371/journal.pcbi.1004286

Chalmers, D. (2000). What is a neural correlate of consciousness ? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions* (pp. 17–39). MIT Press.

Cohen, M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020). Distinguishing the neural correlates of perceptual awareness and postperceptual processing. *Journal of Neuroscience*, *40*(25), 4925–4935. https://doi.org/10.1523/JNEUROSCI.0120-20.2020

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*(263-275), 203.

Crick, F., & Koch, C. (1998). Consciousness and Neuroscience. *Cerebral Cortex*, *8*(2), 97–107.

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*(2), 119–126. https://doi.org/10.1038/nn0203-119

Dehaene, S., & Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

Demertzi, A., Martial, C., Demertzi, A., Tagliazucchi, E., Dehaene, S., Deco, G., Barttfeld, P., Raimondo, F., Martial, C., Fernández-Espejo, D., Rohaut, B., Voss, H. U., Schiff, N. D., Owen, A. M., Laureys, S., Naccache, L., & Sitt, J. D. (2019). Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Science*, *5*(2), 1–12. https://doi.org/10.1126/sciadv.aat7603

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*, 49–59.

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763 LP–13768. https://doi.org/10.1073/pnas.231499798

Fallon, F. (2015). The integrated information theory of consciousness. In J. F. & B. Dowden (Ed.), *The internet encyclopedia of philosophy*. https://iep.utm.edu/int-info/

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T., Galanaud, D., Puybasset, L., Bolgert, F., Sergent, C., Cohen, L., Dehaene, S., & Naccache, L. (2012). Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. *Neuropsychologia*, *50*(3), 403–418. https://doi.org/10.1016/j.neuropsychologia.2011.12.015

Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., Tononi, G., & Pearce, R. A. (2010). Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proceedings of the National Academy of Sciences*, *107*(6), 2681 LP–2686. https://doi.org/10.1073/pnas.0913008107

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, *137*(Pt 10), 2811–2822. https://doi.org/10.1093/brain/awu221

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119–135. https://doi.org/10.1007/s13194-011-0038-2

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., & Hudspeth, A. J. (2012). *Principles of neural science*. McGraw-Hill Education.

Koch, C. (2004). *The quest for consciousness: A neuroscientific approach*. Roberts & Co.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy*, *21*(1), 1–30. https://doi.org/10.3390/e21010017

Michel, M., Beck, D., Block, N., Blumenfeld, H., Brown, R., Carmel, D., Carrasco, M., Chirimuuta, M., Chun, M., Cleeremans, A., Dehaene, S., Fleming, S. M., Frith, C., Haggard, P., He, B. J., Heyes, C., Goodale, M. A., Irvine, L., Kawato, M., … Yoshida, M. (2019). Opportunities and challenges for a maturing science of consciousness. *Nature Human Behaviour*, *3*(2), 104–107. https://doi.org/10.1038/s41562-019-0531-8

Michel, M., Fleming, S. M., Lau, H., Lee, A. L. F., Martinez-Conde, S., Passingham, R. E., Peters, M. A. K., Rahnev, D., Sergent, C., & Liu, K. (2018). An informal internet survey on the current state of consciousness science. *Frontiers in Psychology*, *9*, 2134. https://www.frontiersin.org/article/10.3389/fpsyg.2018.02134

Miracchi, L. (2017). Generative Explanation in Cognitive Science and the Hard Problem of Consciousness. *Philosophical Perspectives*, *31*(1), 267–291. https://doi.org/10.1111/phpe.12095

Morales, J., & Lau, H. (2018). The neural correlates of consciousness. In U. Kriegel (Ed.), *The Oxford handbook of the philosophy of consciousness*. Oxford University Press.

Noel, J.-P., Ishizawa, Y., Patel, S. R., Eskandar, E. N., & Wallace, M. T. (2019). Leveraging non-human primate multisensory neurons and circuits in assessing consciousness theory. *bioRxiv*, 584516. https://doi.org/10.1101/584516

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *The Journal of Neuroscience*, *37*(40), 9593–9602. https://doi.org/10.1101/122267

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, *10*(5). https://doi.org/10.1371/journal.pcbi.1003588

Pautz, A. (2019). What is the integrated information theory of consciousness? A catalogue of questions. *Journal of Consciousness Studies*, *26*(1-2).

Pitts, M. A., Padwal, J., Fennelly, D., Martinez, A., & Hillyard, S. A. (2014). Gamma band activity and the P3 reflect post-perceptual processes, not visual awareness. *NeuroImage*, *101*, 337–350.

Polger, T. W., & Shapiro, L. A. (2016). *The multiple realization book*. Oxford University Press.

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A. G., Bruno, M.-A., Mariotti, M., Boveroux, P., Tononi, G., Laureys, S., & Massimini, M. (2012). Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain*, *135*(4), 1308–1320. https://doi.org/10.1093/brain/awr340

Ruiz de Miras, J., Soler, F., Iglesias-Parro, S., Ibáñez-Molina, A. J., Casali, A. G., Laureys, S., Massimini, M., Esteban, F. J., Navas, J., & Langa, J. A. (2019). Fractal dimension analysis of states of consciousness and unconsciousness using transcranial magnetic stimulation. *Computer Methods and Programs in Biomedicine*, *175*, 129–137. https://doi.org/10.1016/j.cmpb.2019.04.017

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A. G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, *25*(23), 3099–3105. https://doi.org/10.1016/j.cub.2015.10.014

Sasai, S., Boly, M., Mensen, A., & Tononi, G. (2016). Functional split brain in a driving/listening paradigm. *Proceedings of the National Academy of Sciences*, *113*(50), 14444 LP–14449. https://doi.org/10.1073/pnas.1613200113

Silverstein, B. H., Snodgrass, M., Shevrin, H., & Kushwaha, R. (2015). P3b, consciousness, and complex unconscious processing. *Cortex*, *73*, 216–227. https://doi.org/10.1016/j.cortex.2015.09.004.

Tagliazucchi, E., Wegner, F. von, Morzelewski, A., Brodbeck, V., Jahnke, K., & Laufs, H. (2013). Breakdown of long-range temporal dependence in default mode and attention networks during deep sleep. *Proceedings of the National Academy of Sciences*, *110*(38), 15419 LP–15424. https://doi.org/10.1073/pnas.1312848110

Tajima, S., Yanagawa, T., Fujii, N., & Toyoizumi, T. (2015). Untangling brain-wide dynamics in consciousness by cross-embedding. *PLoS Computational Biology*, *11*(11), 1–28. https://doi.org/10.1371/journal.pcbi.1004537

Tononi, G. (2014). *Why scott should stare at a blank wall and reconsider (or, the conscious grid)*. https://www.scottaaronson.com/blog/?p=1823

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, *17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: Here ,There ,and Everywhere? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *31*(1), 12–19. https://doi.org/10.1111/var.12057.Digital

Tononi, G., Sporns, O., & Edelman, G. M. (1999). Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences*, *96*(6), 3257 LP–3262. https://doi.org/10.1073/pnas.96.6.3257

Tsuchiya, N. (2017). "What is it like to be a bat ?" — a pathway to the answer from the integrated information theory. *Philosophy Compass*, *12*, 1–13. https://doi.org/10.1111/phc3.12407

Tsuchiya, N., Taguchi, S., & Saigo, H. (2015). Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience Research*, *107*, 1–7. https://doi.org/10.1016/j.neures.2015.12.007

Voytek, B., Davis, M., Yago, E., Barcelo, F., Vogel, E. K., & Knight, R. T. (2010). Dynamic neuroplasticity after human prefrontal cortex damage. *Neuron*, *68*(3), 401–409. https://doi.org/10.1016/j.neuron.2010.09.018

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II), 8.
https://doi.org/10.33735/phimisci.2020.II.54