Two objections to the IIT theory of consciousness

Abstract

This article discusses two arguments against the IIT theory of consciousness. The first argument says that IIT is wrong in saying that conscious experiences are identical with conceptual structures; they are very different in many ways. The second argument says that the seeming presence of non-conscious integrated information either makes IIT falsified or unfalsifiable. The first argument seeks to show that integrated information is not identical with consciousness; the second argument seeks to show that integrated information is not even always correlated with consciousness.

Introduction

This article challenges two claims made by the Integrated-Information Theory of consciousness (IIT).[1] The first argument says that IIT is wrong in saying that conscious experiences are identical with conceptual structures. I will first explain what a conceptual structure is and what it means to say that it is identical with a conscious experience. Then I will argue that they are not identical since they have many different properties or structural parts.

The second argument says the seeming presence of non-conscious integrated information either makes IIT falsified or makes it unfalsifiable. This argument starts by presenting a case of integrated information which participants in the experiment report as non-conscious but which IIT says must be conscious. The dilemma is that in such a situation either the participants must have a conscious experience of it or else there must be a parallel conscious experience that the participants are not aware of. If the first option is chosen, the problem is that the theory becomes falsified, because the participants have no conscious experience of it. If the second option is chosen, the problem is that the theory becomes unfalsifiable. This is because it predicts outcomes that are seemingly falsified, but the theory is then rescued by appealing to streams of consciousness that are impossible to verify.

If the first argument is right, IIT is wrong in saying that integrated information is *identical* with consciousness, but IIT could still claim that integrated information is always *correlated* with consciousness. If the second argument is right, then IIT is also wrong in saying that integrated information is always correlated with consciousness.

The first argument resembles arguments made by Kelvin McQueen and by Michel and Lau, but I specify differences between our arguments below. The second argument resembles the

[1] In writing this article I have been very much helped by two anonymous reviewers who gave really helpful input to all parts, especially by suggesting many of the sources referenced. Some specific and important contributions by them are referenced at the relevant places. Thanks also to Hedda Hassel Mørch for valuable comments.

unfolding argument by Doerig et al., but I specify below how it is different. The two arguments will now be presented.

Argument number 1: A conscious experience is not identical with a conceptual structure

There are many different problems that a theory of consciousness could try to solve. One can ask what qualia are or what causes qualia, and these are two different questions.[2] One might answer correctly what causes qualia without explaining what qualia are, and one can explain what qualia are without explaining what causes them. Saying that qualia are identical to something else would be an explanation of what qualia are, namely that they are this other thing suggested.

The distinction just made is similar to the distinction made by Michel and Lau between *markers* of consciousness and *constituents* of consciousness. A marker of consciousness indicates that consciousness is present, while the constituents of consciousness are identical with consciousness.[3] Constituents could be contrasted with indicators, causes, correlates, necessary or sufficient conditions, etc.

Michel and Lau use the distinction between markers and constituents of consciousness to distinguish between IIT understood as an empirical theory or a fundamental theory, where the fundamental theory is a theory saying what consciousness is – namely that it is identical with integrated information.[4] The first argument to be presented now is an objection to IIT understood as an explanation of what consciousness is, or fundamental IIT. While Michel and Lau also criticize fundamental IIT, my objection is different from theirs, and I will criticize their objection below.

IIT says that a conscious experience (or a quale) is identical with a conceptual structure.[5] In order to assess this claim, I shall first explain what is meant by a conceptual structure and what it means to say that it is identical with a conscious experience. Then I shall present objections to this claim and answer some possible responses to the objections.

---

[2] In this article, I use "qualia" and "conscious experience" in the same way as Tononi and Koch, as interchangeable terms with the broad sense of any conscious experience – any experience is it like something to have, which disappears in dreamless sleep, but which returns in dreams or waking conditions (C. Koch and G. Tononi, "Christof Koch and Giulio Tononi on Consciousness at the Fqxi Conference 2014 in Vieques," (2014), 20:58.; G. Tononi et al., "Integrated Information Theory: From Consciousness to Its Physical Substrate," Nature Reviews Neuroscience 17 (2016): 450.).

[3] Hakwan Lau and Matthias Michel, "On the Dangers of Conflating Strong and Weak Versions of a Theory of Consciousness," (2019).

[4] Ibid., 3-4.

[5] M. Oizumi, L. Albantakis, and G. Tononi, "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," PLoS Comput Biol 10, no. 5 (2014): 3,14.; Tononi et al., 452-53.

First, what is a conceptual structure?[6] A set of physical elements can be in a certain state where they are related in such a way that the whole system gets a *cause-effect power* on itself and how it will evolve further, independent of extrinsic factors. The obvious example is a brain with neurons, but one could also look at logic gates in a computer. What is important is that the system has internal states that influence each other and are influenced by each other as inputs and outputs.

The cause-effect *power* of the system is its power to cause how it will further evolve; the cause-effect *repertoire* specifies the full cause-effect power of the system *in a particular state*; and the cause-effect *structure* specifies the full cause-effect power of the system with all its subsystems.

The cause-effect structure will have an intrinsic irreducibility, which is a measure of how integrated the information in the system is (measured in units of Φ (big *phi*)). It measures 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).

A maximally irreducible cause-effect structure is called a maximally irreducible conceptual structure (MICS), or just a conceptual structure. One can imagine a cause-effect space where the axes are different possible futures and pasts, and then the conceptual structure will specify how probable the different possible futures and pasts are. The information of the system in a given state is a measure of the extent to which the state constrains its possible past and future states.

Does this mean that according to IIT a conceptual structure is an abstract entity existing in an abstract space of possible futures and pasts? It could seem so, for example when Tononi and others say about physical substrates of consciousness (PSC)," Note that the postulated identity is between an experience and the conceptual structure specified by the PSC, not between an experience and the set of elements in a state constituting the PSC".[7] Kelvin McQueen is also uncertain about how to understand conceptual structures, and asks what their ontology is since they are located in a high-dimensional space.[8]

However, the point seems to be that a conceptual structure should not be identified with the *elements* that constitute the system, but the specific *state* it is in when the internal relations

[6] The following presentation is based on Oizumi, Albantakis, and Tononi. and G. Tononi, "Integrated Information Theory," Scholarpedia 10, no. 1 (2015)..

[7] Tononi et al., 452-53.

[8] Kelvin J. McQueen, "Interpretation-Neutral Integrated Information Theory," Journal of Consciousness Studies 26, no. 1-2 (2019): section 4.1.. McQueen also makes the point that IIT gives no justification for including the possible pasts and futures in their postulates (ibid., section 3.1.).

are a certain way.[9] Tononi specifies that the conceptual structure is physical (see especially the last sentence in this citation):

> It is useful to distinguish between what actually exists – a maximally irreducible cause-effect structure – and its substrate – the set of constituting elements whose state can be observed and manipulated with the available tools in order to reveal the structure. The smallest elements having extrinsic existence, i.e. specifying a cause-effect repertoire, can be considered as the elementary physical substrate out of which everything that exists must be constituted ("atoms" in the original sense of Democritus). Crucially, IIT emphasizes that the substrate does not exist as such, separately from the cause-effect structure it specifies; rather, it exists as that structure. Also, IIT emphasizes that a cause-effect structure is physical, rather than mathematical, because it has its particular cause-effect properties – its particular nature – rather than merely representing a set of numbers that could take other values.[10]

We should thus think of a conceptual structure in a brain as a state which parts of the brain are in at a particular point of time. I understand "being in a state" as meaning that the elements relate to each other in a certain way. Being in that state means that that part of the brain makes certain possible futures and pasts of the same part have certain probability values.

Having now seen what a conceptual structure is, we can move on to the question of what it means to say that a conceptual structure is *identical* with a conscious experience. In philosophy, saying that A is identical with B is commonly taken to mean that A and B refer to one and the same thing, which means that A and B must have all properties or structure parts in common. Leibniz's law expresses this by saying that identicals are indiscernible.

The concept of identity that interests us here is not identity over time, where A last year may be identical with B this year, even if A and B have different properties, but rather identity at a given point of time. One can distinguish between type and token identity at a point of time, where two different objects can be type identical in the sense of having the same internal structure, and nevertheless not be token identical since they may have different external relations and are located at different places.

IIT seems to use the concept of identity in the strict sense of token identity at a given point of time. Tononi et al. says that "An experience is identical to a conceptual structure, meaning that every property of the experience must correspond to a property of the conceptual structure and vice versa."[11] The term "correspond" is still open to interpretation, but another place Tononi writes that "The overall 'form' of the conceptual structure or *quale sensu lato*

[9] Thanks to an anonymous referee for helping me see clearly how to interpret the concept of conceptual structures.

[10] Tononi.. But it is not entirely clear, since in the footnote after this quote (footnote 28) Tononi writes: "It is intriguing to consider to what extent the physical world has *intrinsic existence* (cause-effect power from its own perspective—in and of itself) in addition to *extrinsic existence* (cause-effect power from the perspective of an observer who can perform interventions on it and sample the results)."

[11] Tononi et al., 452-53.

(Q) (constellation of stars) is identical (≡) to the quality of the experience, how the experience feels."[12]

In the last quote, the triple bar (≡) is used to specify the meaning of "identical". Unfortunately, the triple bar has different meanings in different contexts, and it is not defined further by Tononi. But one interpretation of the triple bar, which is a plausible interpretation of Tononi, is that it means that that the terms refer to one and the same thing.

I shall interpret "being identical with" in the strict sense that it means sharing all properties, since this is what is required for IIT to explain what qualia is. However, below I shall consider an objection saying that conceptual structures and qualia have different properties since one is the extrinsic perspective and the other is the intrinsic perspective of the same structure.

We have now seen what a conceptual structure is and what it means to say that it is identical with qualia, qualia being used in the broad sense of referring to any conscious experience which it is like something to have. In the following, I shall raise some objections to this claim.

If qualia and conceptual structures are identical in the strict sense, every property should be identical and they should have every part of their structure in common. Instead they seem to have very many different properties and structure parts.[13] On the one hand you have the conscious experiences and their properties. The structure or properties of conscious experiences of color is that they come in degrees of hue, brightness and saturation. The structure or properties of conscious experiences of sound is that they come in degrees of pitch, loudness and timbre. The structure or properties of conscious experiences of taste is that they come in degrees and combinations of salt, sweet, bitter, sour and umami, but this is also very influenced by odor. And so it continues, and odors, feelings and thoughts are far more complex than the examples already given.

On the other hand, you have conceptual structures. In the case of human consciousness, the relevant conceptual structures consist of neurons related in such a way that their action potentials and firing patterns at the given point of time make different possible pasts and different possible futures of that system have certain probabilities. The system being in a state with these probability values are the constellations that are said to be identical with qualia. Maybe they are just a way to express other properties in the relations between neurons that are better thought of as identical with qualia. In any case, these properties seem completely

[12] Tononi.

[13] "Properties" and "structure parts" refer to the same here. Substance ontologists prefer to speak of the properties of substances, while structuralist ontologists reject substance and reinterpret properties as structure parts. I do not enter this discussion here, but use both terms to communicate with both sides.

different from the properties of qualia. There is no reason to think that they are the same – and thus identical.

Instead of being identical, the properties of qualia and conceptual structures seem very different and inconsistent with each other. The conceptual structure is composed of neurons which again are composed of elementary particles related to each other in a certain way at a certain location in spacetime (thus it is physical), while many qualia show no sign of being located in spacetime or made of elementary particles at all. Everything about a conceptual structure can be seen and measured from a third-person perspective, while qualia are only accessible from a first-person perspective.

To say that qualia and a physical structure are identical when their properties seem so different is of little value if it is merely a claim without support showing why qualia are nevertheless physical when they seem to lack common physical properties. If there is much more evidence of differences than of identity, one should conclude that they are different and not identical. This objection applies regardless of whether a conceptual structure is interpreted physically or non-physically, since the properties of the conceptual structure as it is described in IIT is in both cases different from the properties of qualia.

Further, why should we think that the cause-effect power of a physical system to influence how it evolves in the future should have any connection to qualia? There are many physical systems that have a cause-effect power to influence how it evolves in the future, where there seems to be no reason to think that consciousness is present. An article by Oliver, Seddon and Trask lists plenty of self-forming materials, organs, organisms and mechanisms in nature that do this, such as liquid crystal, the pulvinus, hydrogels, etc.[14] It seems plausible to assume that there are many internally complex structures in nature which in different states have different probability values for different futures and pasts, without there being any indication of consciousness present. IIT says that more is required for consciousness, but why think there is any connection between consciousness and the capacity for influencing possible futures at all?

Here are some possible responses that an IIT-defender could offer to these objections:

Firstly, one could respond that it is just a brute fact that there is a metaphysical connection between these properties, and one may say that while we do not see how the structures are the same today, it is worthwhile to try to discover this in the future. This is certainly possible, but we should then say that today we have no good reason to think that conceptual structures and qualia are identical even if the future may prove us wrong (which the future always can in every question).

[14] Kate Oliver, Annela Seddon, and Richard S. Trask, "Morphing in Nature and Beyond: A Review of Natural and Synthetic Shape-Changing Materials and Mechanisms," Journal of Materials Science 51, no. 24 (2016).

Secondly, one could respond that the descriptions are very different since they are being described in different theoretical frameworks or at different levels of descriptions. This is similar to how John Searle has defended physical descriptions of qualia by arguing that you can describe a motor very differently at different ontological levels, even if it is one and the same motor.[15]

My response to this is to say that if such is the case, one needs to show how to translate from one description to the other. In the case of qualia and conceptual structures, they are described with seemingly inconsistent properties, and one should then face the challenge of showing why they are nevertheless the same.

A third response, similar to the previous one, could be to say that while qualia and conceptual structures are different, they are different aspects of one and the same structure, where the difference is between how they are intrinsically and extrinsically – or from an internal or external perspective. We can describe extrinsically a physical relation between neurons and represent it abstractly as a conceptual structure in concept space, but intrinsically the physical system with this specific structure experiences its structure as qualia. This could be a response offered by Oizumi et al., when they say that "an experience is thus an *intrinsic property* of a complex of mechanisms in a state"[16], and by "intrinsic" they add throughout the article that it means "from the system's own perspective". It could also be a possibility Tononi contemplates when he first says that conceptual structures are physical, but then adds in a footnote:

> It is intriguing to consider to what extent the physical world has *intrinsic existence* (cause-effect power from its own perspective—in and of itself) in addition to *extrinsic existence* (cause-effect power from the perspective of an observer who can perform interventions on it and sample the results).[17]

The problem with this response is that it is obscure what intrinsic property or intrinsic existence from its own perspective means. Presumably everything has been described of the system and its structure, so if it has other properties, something must account for that. What is the "inside" or "the self" that can have an own perspective or add extra properties "intrinsically"? Kelvin McQueen makes the same point that it is obscure to add "intrinsic" or "inside" to the explanation.[18]

Intrinsicality can be defined in different ways. Since I am a structuralist, I believe that nothing is more than internal structure and external relations, all of which can be described. If intrinsicality means internal structure, this should be describable. Intrinsicality may mean an

[15] John R. Searle, Consciousness and Language (New York: Cambridge University Press, 2002), 27.

[16] Oizumi, Albantakis, and Tononi, 3., emphasis in original.

[17] Tononi. n28.

[18] McQueen, section 4.1.

irreducible structure which cannot be further explained or defined, and all theories include some irreducible parts. But if consciousness is something irreducible, it is something different than being reducible to physical conceptual structures. While the discussions of structuralism and intrinsicality cannot be made here, a development of what intrinsicality means should be offered by IIT to make sense of consciousness as an intrinsic property.

A fourth response could be to accept that "identity" is the wrong word to use, but see if the relation can be made more precise.[19] Tsuchiya et al. have suggested category theory as a way of making the relation precise, since this is a theory that can be used to describe degrees of qualitative similarity (from existence of a functor relation, to existence of adjunction, to the entities being categorically equivalent, then categorically isomorphic, and finally being identical).[20]

In category theory, a category is defined as a set of objects and arrows, and everything that exists is either an object or an arrow relating objects. Tsuchiya et al. find that conceptual structures are clearly categories, and argue that qualia can also be understood as objects with arrows. They argue that there is a functor relation between conceptual structures and qualia, since the functor relation is extremely flexible, and that IIT should try to use category theory to map more precisely the relation between conceptual structures and qualia.[21]

It seems to me less fruitful to measure the degree of qualitative similarity than to just discuss as precisely as one can which structures are identical and which structures are different. I find it quite vague and un-enlightening to think of qualia (like existential anxiety) as objects with arrows that can be related to something physical with an extremely flexible functor relation. But I do support the idea of trying to map as closely as one can the properties of conceptual structures and qualia to see what one can learn about its correlations. I just think that so far we have no good reason to suggest that it is a relation of identity.

A fifth response would be to say that IIT is the best theory we have of consciousness, with many interesting indications of a close relation between integrated information and consciousness. Even if correlation does not imply identity, the correlation would have a neat explanation if it did turn out to be identity, and thus this could be the research program or hypothesis of IIT even if they have not yet been able to show in any convincing way that there is an identity.

[19] Thanks to an anonymous referee for suggesting this response and the article by Tsuchiya et al.

[20] Naotsugu Tsuchiya, Shigeru Taguchi, and Hayato Saigo, "Using Category Theory to Assess the Relationship between Consciousness and Integrated Information Theory," Neuroscience Research 107, no. June (2016).

[21] Ibid.

I have no problem with investigating interesting hypotheses either. I just think that IIT should say that as of today we have no reason to think that there is an identity as opposed to merely a correlation, and above I have offered reasons to think that it is not an identity, because the properties seem inconsistent. In the next section, I shall problematize how strong the correlation is as well. I believe that integrated information is an important part of the work that the brain does and can do consciously, but that this work can be done non-consciously as well, which means that integrated information is necessary for making informed decisions but not for being conscious, even if informed decisions can be conscious. If this is right, it explains why integrated information is often correlated with consciousness, but not always.

Before turning to the second objection, I shall end this section with a comparison between the objection I have made here and a similar objection made by Michel and Lau. According to Michel and Lau, IIT fails as a fundamental explanation of what consciousness is. They argue that in order to show that consciousness is identical with some neural correlates of consciousness (NCC), you need to demonstrate that they are necessary for the presence of consciousness. As a comparison they argue that you can show that water is identical with H_2O by showing that the presence of H_2O is necessary for there to be water. Since IIT fails to show this, Michel and Lau argues that integrated information is merely a marker of consciousness and not a constituent.[22]

If water is identical with H_2O only, the presence of H_2O is necessary for the presence of water. But what we think of as one and the same substance (water) could turn out to be different chemical substances, such that the presence of one of these is sufficient for the presence of what we think of as water, while none of them are actually necessary. In fact, there are many different chemical combinations that give us what we think of as water (only chemists can tell the difference), such as $H_2{}^{17}O$, $H_2{}^{18}O$, $HD^{16}O$, $D_2{}^{17}O$, and $T_2{}^{18}O$.[23]

Maybe what we experience as consciousness comes in different forms that are identical with different physical structures. But if someone is able to show that what we know as consciousness have all properties in common with some X, they may well say that (our form of) consciousness is identical with this X – in the sense of having all properties in common – even if the presence of X is not necessary for all kinds of consciousness, since possibly the presence of Y could give a similar kind of consciousness.

When it was discovered that water is H_2O, we could see that (what we think of as) water has the same properties as H_2O at a coarse-grained level: relating to each other in a fluid-like

[22] Lau and Michel.

[23] Michael Weisberg, "Water Is Not H20," in Philosophy of Chemistry: Synthesis of a New Discipline, ed. Davis Baird, Eric Scerri, and Lee McIntyre, Boston Studies in the Philosophy of Science (Doordrecht: Springer, 2006).. What we call H_2O is more precisely $H_2{}^{16}O$

way, being transparent, heatable, freezable, etc. Later we have learned that at a more fine-grained level, there are different forms of water identical with different chemical compounds.

While Michel and Lau require that IIT should show that consciousness and NCC are identical at a fine-grained level and argue that IIT fails to do this, I only require IIT to show that consciousness and NCC have the same properties at a coarse-grained level. Yet I argue that IIT fails to do this. While Michel and Lau also require that IIT should show that a NCC is necessary for claiming identity with consciousness and argue that IIT fails to do this, I only require that they should show that the properties are the same. Again, I argue that IIT fails to do this. Michel and Lau have a very demanding requirement, whereas I argue that IIT fails a much less demanding requirement.

Argument number 2: A dilemma for IIT

The second argument challenges the claim that integrated information is identical with conscious experiences and even that it is necessarily correlated with conscious experiences. The argument takes form as a dilemma, where both horns of the dilemma appear to be unacceptable for IIT.[24]

The starting point for the dilemma is the following claim: There seems to be integrated information in the brain that is not conscious. For example Damasio and colleagues performed an experiment where people were asked to draw cards from various decks. Some decks were good, leading to a reward, and some were bad, leading to a punishment. There was also a system determining which decks were good and which were bad, so that if you cracked the code you could just draw good cards. The subjects played the game while their skin conductance was measured. The interesting thing was that it seemed that the code was cracked non-consciously several minutes before the players understood it consciously and started drawing only winning cards. After a while, they would get one type of skin response just before drawing from every bad deck and another type of skin response just before drawing from every good deck. This was so consistent that somehow some part of the brain must have cracked the code. Yet the person could not consciously tell this and would keep drawing bad cards.[25]

This process seems very similar to the kinds of processes that can be conscious, and Damasio presents several cases of non-conscious thinking, feeling, remembering, etc.[26] It is hard to see

[24] I am grateful to an anonymous reviewer for the suggestion of formulating this argument as a dilemma this way.
[25] Antonio R. Damasio, Self Comes to Mind: Constructing the Conscious Brain (New York: Pantheon Books, 2010), 276.
[26] Ibid.

how this could come about with no integration of information, since much simpler systems (like thermostats) are said to integrate information.

Given that there is integrated information in the brain in examples like the one above, this can be used to formulate a dilemma for IIT. According to IIT, if there is integrated information, there is consciousness present. This consciousness would then either have to be part of the conscious experience of the participants in the card experiment, or it would not. If it is not part of their conscious experience but nevertheless is conscious (as IIT implies), then the participants in the card experiment must have a parallel stream of consciousness that they are not aware of.

The dilemma is then the following: Given the presence of integrated information in experiments like the card experiment, and given that integrated information implies consciousness, there must either be a conscious experience of it had by the participants or a parallel conscious experience that the participants are not aware of. If the first option is chosen, the problem is that the theory becomes falsified, because the participants have no conscious experience of cracking the code. If the second option is chosen, the problem is that the theory becomes unfalsifiable because it predicts outcomes that are seemingly falsified, but the theory is rescued by appealing to streams of consciousness that are impossible to verify.

I guess it would be possible to reject the dilemma by rejecting that there is integrated information present in the first place. However, the dilemma can be strengthened.[27] Kelvin McQueen has argued that while activity in the cerebellum is considered non-conscious, there are pockets within the cerebellum with maximal (big) phi spikes that should be conscious according to IIT, but which are never reported by anyone as conscious experiences.[28]

The dilemma can then be restated: Should we think of these cerebellum pocket activities as conscious or not? Again: If the first option is chosen, the problem is that the theory becomes falsified, because we have no conscious experience of conscious activity in the cerebellum. If the second option is chosen, the problem is that the theory becomes unfalsifiable. It predicts outcomes that are seemingly falsified, but is rescued by appealing to impossible to verify streams of consciousness.

A possible response by IIT to these dilemmas could be to say that while it may be unfalsifiable that parallel streams of consciousness could exist, we could in fact design experiments to give support to the existence of parallel streams of consciousness. This strategy is in fact explored by Sasai and others (including Tononi). In one experiment, drivers

[27] Thanks to an anonymous referee for this suggestion.

[28] K. J. McQueen, "Illusionist Integrated Information Theory," Journal of Consciousness Studies 26, no. 5-6 (2019): 150.

are given parallel tasks with their brains monitored, and the question is whether there in such cases are a functional split in the brain similar to anatomical splits.[29]

The results show split attention, but the result is compatible with either one single conscious stream switching between tasks, one of the tasks is done non-consciously, or there are two parallel conscious streams. Sasai et al. find that there is slight support for two parallel streams because their task performance matches the results in split brain patients.[30]

Whether this result supports parallel streams of consciousness depends on whether there are parallel streams of consciousness in split brain patients. The authors say that in split brain patients "two separate streams of consciousness coexist within a single brain, one per hemisphere",[31] with reference to two articles by Michael Gazzaniga. One of them is entitled "When the cerebrum is divided surgically, it is as if the cranium contained two separate spheres of consciousness".[32] The other article makes the same kind of claim with the term "as if", and ends by asking what it means to split a brain.[33]

Research on split brains clearly shows that the brain can support streams of consciousness with different and disconnected contents, but they do not show that there are two streams of consciousness existing at the same time caused by one brain. We should thus distinguish there being two *parallel* streams of consciousness and one stream of consciousness with *disconnected* or *switching* content.

Tim Bayne has argued well that split-brain patients do not have two streams of consciousness, but that we should interpret their experiences instead in light of a switch model.[34] We know from experiments that the content of our consciousness can switch instantaneously. An example is binocular rivalry, where inconsistent input is given to one eye each and one is conscious for a while at one input, then switches to another. This known model explains the findings from split-brain patients and how there is some integration from both hemispheres in such patients.[35]

---

[29] Shuntaro Sasai et al., "Functional Split Brain in a Driving/Listening Paradigm," Proceedings of the National Academy of Sciences 113, no. 50 (2016).
[30] Ibid., 14447-48.
[31] Ibid., 14444.
[32] Ibid., 14449., referring to Michael S. Gazzaniga, "The Human Brain Is Actually Two Brains, Each Capable of Advanced Mental Functions. When the Cerebrum Is Divided Surgically, It Is as If the Cranium Contained Two Separate Spheres of Consciousness.," Scientific American 217, no. 2 (1967).
[33] "The Split-Brain: Rooting Consciousness in Biology," Proceedings of the National Academy of Sciences 111, no. 51 (2014).
[34] Tim Bayne, "The Unity of Consciousness and the Split-Brain Syndrome," Journal of Philosophy 105, no. 6 (2008).
[35] Ibid., 286, 94-99.

The way that this second objection challenges IIT with falsification or unfalsifiability has some resemblance to the unfolding argument made by Doerig et al. Their point is that IIT acknowledges that feed-forward networks (with zero phi) can do anything that recurrent networks (with phi > 0) can do. It then seems that either IIT must accept that feed-forward networks can be conscious (falsifying their own theory) or say that feed-forward networks are not conscious even if they are empirically indistinguishable from recurrent networks (which makes the theory unfalsifiable).[36]

The Doerig dilemma would arise if one responds that there is no integrated information in the card game example, but rather a feed-forward network with zero phi. If one acknowledges that there is integrated information in the card game example, the dilemma becomes that either it should be experienced as conscious (which it is not, making the theory falsified) or it is unexperienceable for the participants (making the theory unfalsifiable).

To sum up, the seeming presence of non-conscious integrated information either makes IIT falsified or unfalsifiable, which in either case makes it problematic to argue that consciousness is identical with integrated information or even that it is always correlated.

Conclusion

In this article I have discussed two arguments. The first argument says that IIT is wrong in identifying conscious experiences with conceptual structures since they are different in many ways. The second argument says that the seeming presence of non-conscious integrated information either makes IIT falsified or unfalsifiable, implying both that integrated information is not identical with conscious experiences and that it is not always correlated either.

Sources

Bayne, Tim. "The Unity of Consciousness and the Split-Brain Syndrome." *Journal of Philosophy* 105, no. 6 (2008): 277-300.

Damasio, Antonio R. *Self Comes to Mind: Constructing the Conscious Brain*. New York: Pantheon Books, 2010.

Doerig, Adrien, Aaron Schurger, Kathryn Hess, and Michael H. Herzog. "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness." *Consciousness and Cognition* 72 (2019): 49-59.

Gazzaniga, Michael S. "The Human Brain Is Actually Two Brains, Each Capable of Advanced Mental Functions. When the Cerebrum Is Divided Surgically, It Is as If the Cranium Contained Two Separate Spheres of Consciousness.". *Scientific American* 217, no. 2 (1967): 24-29.

———. "The Split-Brain: Rooting Consciousness in Biology." *Proceedings of the National Academy of Sciences* 111, no. 51 (2014): 18093–94.

[36] Adrien Doerig et al., "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness," Consciousness and Cognition 72 (2019): 53.

Koch, C., and G. Tononi. "Christof Koch and Giulio Tononi on Consciousness at the Fqxi Conference 2014 in Vieques." 2014.
Lau, Hakwan, and Matthias Michel. "On the Dangers of Conflating Strong and Weak Versions of a Theory of Consciousness." (2019). Published electronically June 2019. doi:10.31234/osf.io/hjp3s.
McQueen, K. J. "Illusionist Integrated Information Theory." *Journal of Consciousness Studies* 26, no. 5-6 (// 2019): 141-69.
McQueen, Kelvin J. "Interpretation-Neutral Integrated Information Theory." *Journal of Consciousness Studies* 26, no. 1-2 (2019): 76-106.
Oizumi, M., L. Albantakis, and G. Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." *PLoS Comput Biol* 10, no. 5 (2014).
Oliver, Kate, Annela Seddon, and Richard S. Trask. "Morphing in Nature and Beyond: A Review of Natural and Synthetic Shape-Changing Materials and Mechanisms." *Journal of Materials Science* 51, no. 24 (December 01 2016): 10663-89.
Sasai, Shuntaro, Melanie Boly, Armand Mensen, and Giulio Tononi. "Functional Split Brain in a Driving/Listening Paradigm." *Proceedings of the National Academy of Sciences* 113, no. 50 (2016): 14444-49.
Searle, John R. *Consciousness and Language*. New York: Cambridge University Press, 2002.
Tononi, G. "Integrated Information Theory." *Scholarpedia* 10, no. 1 (2015): 4164.
Tononi, G., M. Boly, M. Massimini, and C. Koch. "Integrated Information Theory: From Consciousness to Its Physical Substrate." *Nature Reviews Neuroscience* 17 (2016): 450-61.
Tsuchiya, Naotsugu , Shigeru Taguchi, and Hayato Saigo. "Using Category Theory to Assess the Relationship between Consciousness and Integrated Information Theory." *Neuroscience Research* 107, no. June (2016): 1-7.
Weisberg, Michael. "Water Is *Not* H20." In *Philosophy of Chemistry: Synthesis of a New Discipline*, edited by Davis Baird, Eric Scerri and Lee McIntyre. Boston Studies in the Philosophy of Science, 337-45. Doordrecht: Springer, 2006.