Adam Pautz

What is the Integrated Information Theory of Consciousness?

A Catalogue of Questions

Abstract: *In this paper, my goal is modest. I will not argue that the integrated information theory (IIT) is false. Instead, I will raise a number of basic questions about what the theory is. As long as proponents of IIT do not address these questions, they have not put a clear theory on the table that can be evaluated as true or false.*

Integrated information theory (IIT) holds that the 'level' or 'amount' of consciousness in a system is determined by the amount of 'integrated information' (Φ) in the system. Roughly, this is 'the amount of information generated by a complex of elements, above and beyond the information generated by its parts' (Tononi, 2008, p. 216). The theory promises to provide the holy grail: a mathematically precise and elegant theory linking conscious experiences to underlying physical states.

The theory itself is neutral on the mind–body problem. In particular, it is neutral between 'emergentism' and 'reductionism' about consciousness. Emergentists could take the principles of IIT to be fundamental psychophysical laws linking physical states of integrated information with the emergence of distinct states of consciousness. Reductionists could take them to be underwritten by psychophysical identities between states of consciousness and physical states of integrated information. IIT provides something that both reductionists

Correspondence:
Adam Pautz, Brown University, Providence, RI 02912, USA.
Email: adam.pautz@gmail.com

Journal of Consciousness Studies, **26**, No. 1–2, 2019, pp. 188–215

and emergentists need: a simple and elegant theory linking consciousness to the physical world.[1]

Many find IIT attractive. In fact, Christof Koch has recently declared that 'it's the only really promising fundamental theory of consciousness' (quoted in Zimmer, 2010). But it also faces problems. In his discussion, 'Why I Am Not an Integrated Information Theorist', Scott Aaronson has argued that it implies that consciousness is present in some very simple physical systems, because simple physical systems may have arbitrarily high levels of integrated information (Φ). For instance, he describes a simple 'Vandermonde contraption' and argues that IIT 'predicts that this Vandermonde contraption would be *billions of times more* conscious than you are' and indeed that such contraptions 'can be *unboundedly more* conscious than humans are' (Aaronson, 2014, my italics). Critics like Aaronson have regarded these consequences as amounting to a *reductio ad absurdum*. By contrast, proponents see them as interesting discoveries (Tononi, 2014; Tononi and Koch, 2015).

In my view, such standard objections to IIT are not decisive. It is true that IIT has strange predictions. But if a theory of consciousness fits the data from humans and is more elegant than the alternatives, maybe we should accept the theory even if it has some strange predictions regarding the consciousness of non-humans. After all, some of our best physical theories have strange predictions too. The problem of consciousness is hard. We should keep an open mind.

My primary aim is a modest one. I am not after a refutation of IIT. Instead, my main aim is just to raise a series of interpretive questions. I invite enthusiasts of IIT to say more about what they want to explain and how their approach is meant to explain it.

---

1 By *reductionism*, I mean the view that states of consciousness are *identical* with complex physical or functional states. By *emergentism*, I mean the view that there are no such identities and that states of consciousness are dependent on physical or functional states by way of special 'nomological laws' or 'grounding laws'. As Fodor put it, 'maybe the hard problem shows that not all basic laws are laws of physics… some of them are laws of emergence' (2007). Examples of emergentism include dualism (Chalmers, 1995) and emergent physicalism (Rosen, 2010). Tononi appears to favour a reductive form of integrated information theory (Oizumi, Albantakis and Tononi, 2014, p. 3). Koch, a recent convert to integrated information theory, is harder to classify. He says he is a 'romantic reductionist'. Yet he also says that he is dualist who thinks that consciousness 'is something fundamentally different from the material thing causing it and that it can never be fully reduced to the physical properties of the brain' (2012, p. 119).

In particular, I will address two issues. In the first instance, IIT is formulated as a theory of the physical basis of the 'level' or 'amount' of consciousness in a system. In addition, integrated information theorists have tried to provide a systematic theory of how physical states determine the *specific qualitative contents* of episodes of consciousness: for instance, an experience as of *a red and round thing* rather than *a green and square thing*. First, I will raise a series of questions about the central explanatory target, the 'level' or 'amount' of consciousness (§1). Second, I will raise some questions about the explanation of qualitative content (§2).

1. What Do Integrated Information Theorists Mean by 'Level of Consciousness'?

Typically, one thing theories of the physical basis of consciousness attempt to provide is an answer to the following question:

> *The Ignition Question*: When is a physical system conscious? What is the correct, modally robust principle that tells us whether or not a system is conscious, given its physical state?

Different theories provide different principles: for instance, functional theories typically appeal to a condition involving accessibility for global control (in reasoning, action, etc.), whereas biological theories require the presence of a specific type of neural state.[2]

[2] For biological theories of consciousness, see Lamme (2006; 2010) and Block (2019). For accessibility theories, see Baars (1988), Dehaene (2014), and Tye (2000). Emergentists about consciousness could accept or co-opt these theories but take them to be non-reductive. For instance, Chalmers (forthcoming) suggests that dualists may posit a contingent psychophysical law connecting consciousness with 'cognitive accessibility' somehow spelled out in physical terms. (A difficulty for this combination of views is that, since it is indeterminate at what moment in evolutionary history the states of organism first satisfied the very vague 'cognitive accessibility' condition (Tye, 2000, chapter 8), it implies that it is metaphysically indeterminate exactly when novel, emergent conscious states first appeared in the world. It is hard to make sense of such metaphysical indeterminacy.) Tononi and Koch (2015, p. 2) consider very crude accessibility theories based on *actual verbal report* and raise counter-examples involving dreams and animals; but these are not problems for sophisticated accessibility theories since the states of such individuals are poised to influence beliefs and guide behaviour even if they do not cause verbal reports. In fact, there can be no decisive counter-examples to such theories because where there is absolutely no accessibility whatever, as for instance in Block's G.K. example (2007, p. 498), we can have at best very indirect and weak reason to think that there is consciousness (Chalmers, 1998/ 2010, p. 99, fn. 2).

IIT departs from standard theories in that it is primarily a theory of the 'level' or 'amount' of consciousness in a system, not just a theory of when consciousness is present or absent. Here are some representative statements:

> The IIT claims that consciousness is not an all-or-none property, but is graded. (Tononi, 2008, p. 236)
>
> The quantity of consciousness corresponds to the amount of integrated information [Φ] generated by a complex of elements. (*ibid.*, p. 216)
>
> IIT postulates that the amount of integrated information [Φ] that an entity possesses corresponds to its level of consciousness. (Koch, 2009)
>
> The quantity or level of consciousness [in a system] is measured by its [Φ] value. (Tononi and Koch, 2015, p. 9)

Thus, IIT is framed with the observational term 'the level or amount of consciousness' and the theoretical term 'Φ'. (Compare 'level of heat is determined by mean molecular kinetic energy'.) Therefore, to understand what is being claimed, we must understand both of these terms.

Here I will simply assume that Φ is a clear measure. Roughly, it is 'the amount of information generated by a complex of elements, above and beyond the information generated by its parts' (Tononi, 2008, p. 216). There are multiple ways of defining 'Φ'. In a recent development of the theory (integrated information theory 3.0), Oizumi, Albantakis and Tononi (2014) supply the currently favoured measure. For systems as complex as the human brain, it may be impossible in practice to get accurate estimates of the level of Φ.[3] Still, I will assume for the sake of discussion that Φ is a precise measure.

Instead, I will develop a basic question about the key observational term, 'the level or amount of consciousness'. To take a hypothetical example: suppose that we have determined that the Φ-value of a person's brain has decreased due to intoxication (say, it has halved). Then IIT predicts that her 'level or amount of consciousness' has

[3] For this reason, there is a real question of whether IIT is empirically testable. Casali *et al.* (2013) have found that another, more tractable index, PCI (Perturbational Complexity Index), correlates with consciousness. And they suggest that PCI in turn correlates with level of integrated information Φ. But Sitt *et al.* (2013, p. 552) object that 'the PCI is not directly related to Tononi's [Φ] measure proposed by IIT as a marker of consciousness'. So it is possible to question whether there is at present any reason at all to believe IIT and its radical implication that simple 2D gates are conscious.

decreased (it has halved). If IIT is to be a meaningful and testable theory, we must have an independent grasp on what this kind of prediction means, and we must be able to confirm it by observation.

However, it is not at all immediately clear what 'level of consciousness' means. While we might all easily know what 'consciousness' means, '*level* of consciousness' stands in need of clarification. For our conscious experiences are graded along *multiple* dimensions. Which one do integrated information theorists have in mind?[4]

Let me begin by listing some of the dimensions along which conscious experiences are graded. Then I will be able to formulate my question more exactly.

A preliminary: I will make the standard assumption that episodes of consciousness are essentially *intentional*. They have built-in intentional contents, which I will call *qualitative contents*. These contents can be specified by a *proposition*. Since the content of a typical experience is very rich, we cannot fully specify the content of an experience in language. At best, we can give an approximation. For instance, the built-in qualitative content of an experience might be something like: *there is a thing with a specific shade of red and a somewhat bulgy, tomato-shape directly in front of me*. In an hallucination case, the content does not correspond to reality: there is no reddish or round thing in the world or in the brain.[5]

Here now is a partial list of some of the dimensions along which experiences are graded:

4 See Pautz (2015) for the point that conscious states vary along multiple dimensions, so that talk of 'the level or amount of consciousness' is very unclear. Bayne, Hohwy and Owen (2016) develop the same basic point. Here I am drawing out a corollary of this point: IIT, because it is a theory of 'the level or amount of consciousness' in a system, is likewise fundamentally unclear.

5 For discussions of the intentionality of experience, see for instance Tye (2000) and Chalmers (2006). By the 'intentionality of experience' I just mean that having a tomato-like experience, for instance, essentially involves the seeming-presence of a *round* thing; it is essentially an experience as of a *round* thing. This is a pre-theoretical claim framed in ordinary language. It seems to be an obviously correct description of the phenomenology of experience. It follows that a full characterization of the essence of the experience must mention the feature *round* that enters into the content of the experience. Some have denied this. For instance, Papineau (2016) has recently defended the view that to have the tomato-like experience is to undergo an internal neural state whose essential nature be can fully described in terms of *types of neurons* and the *times*, *directions*, and *intensities* at which they fire, without using the spatial term *round* at all. But this goes against the phenomenology.

Intensity Level. Conscious experiences differ in the intensity of their qualitative contents. For instance, consider an experience of a noise at the threshold of hearing versus the experience of a rock concert, or the experience of a dull colour versus the experience of a bright colour.

Complexity Level. Individual conscious experiences differ in the complexity of their contents. For instance, an olfactory experience of a minty smell has a relatively simple content, along the lines of *something minty hereabouts*. By contrast, a visual experience of a street scene in Paris has a very complex or 'rich' content involving the attribution of numerous properties (shapes, colours) to many objects and regions. Indeed, some think that the content involves the attribution of high-level properties like *being a happy face*, *being a car*, and so on. Maybe the contents of the experiences of primitive creatures like snails are much simpler than the contents of our experiences.

Determinacy Level. Conscious experiences differ in the *determinacy* of their content. If you look at a tomato right in front of you, the content of your experience specifies a more-or-less determinate colour, and a more-or-less determinate shape. If you move the tomato to the periphery of your visual field, then the spatial content becomes more 'degraded', less determinate and more 'determinable', because of the lower spatial resolution of peripheral vision. Likewise, stimuli presented in low-contrast conditions might be experienced with little precision. It is natural to think that the contents of experiences in dreams and imagery are indeterminate and degraded. And maybe this is something that happens if a person becomes highly sedated. Determinacy Level is not the same as Complexity Level: if you look at a blank wall in front of you, the content of your experience has a low complexity level but it has a high determinacy level: it specifies a more-or-less determinate colour.

Access Level. So far, we have seen that the contents of our experiences vary along a few dimensions. There may also be variation in our cognitive and functional *access* to the qualitative contents of our experiences. For instance, maybe there is variation in cognitive-uptake: 'how much' of the qualitative information represented by experience is actually taken up into working memory for the control of action. On some views, the access level of an experience could in principle be zero. For instance, Ned Block speculates that the patient G.K. might have an isolated

experience with a determinate content specifying the presence of a face, but where this content is not and indeed cannot be cognitively accessed at all (2007, p. 498). At the other extreme, if you have an experience with a very simple qualitative content, and fully attend to that content, you may be able to completely cognitively 'take in' precisely that content (say, believe it) while you are having the experience. This would be the upper bound of access level.

Richness of Experiential Repertoire. So far, we have seen that *specific experiences* can be roughly ordered along various dimensions. We can also order creatures along a dimension that we might call 'experiential repertoire'. This I intend to be a capacity notion. Roughly, by the 'experiential repertoire' of a creature I mean the 'number' of distinct experiences (types and tokens) the creature has the capacity to have. For instance, an ordinary adult human has colour experiences, auditory experiences, and so on. She can have endless distinct visual experiences presenting different combinations of colours and shapes. This defines her experiential repertoire. If she suddenly became blind, then there would be a reduction in her experiential repertoire. Newborns have a less rich experiential repertoire than adults. Maybe snails have a still smaller experiential repertoire. In that sense, snails have a 'lower level of creature-consciousness' than humans.

Having distinguished these dimensions along which conscious states are graded, I can reformulate my question in this way:

> **Q1:** By 'level of consciousness', do integrated information theorists mean to refer to one of the dimensions just described? Or do they mean to refer to some other dimension not yet specified? Exactly *what* aspect of consciousness is it that, on their theory, is supposed to wax and wane precisely in proportion to Φ?

As long as integrated information theorists do not address the question, we literally do not know what they are claiming and their theory is untestable. It is as if they are saying 'the bling-value of a system's consciousness is measured by the Φ-value of that system', without giving any indication of what they mean by 'bling-value'.

I also have some more specific questions. For any dimension or scale, we can ask some basic questions. To begin with, we can ask whether it forms a *ratio scale* or a mere *ordinal scale* (Gescheider,

1997). And we can ask if it has an upper bound or it is unbounded. So the following questions arise:

> **Q2:** Does whatever dimension that integrated information theorists are referring to by 'level of consciousness' have ratio scale?
> **Q3:** Does the dimension they have in mind have an upper bound or is it unbounded?

How might integrated information theorists answer Q1–Q3?

Let me start with Q2 and Q3. Integrated information theorists do not explicitly address these basic questions. But I think it is clear that, whatever they mean by 'level of consciousness', their theory straightforwardly implies that it has a ratio scale and is unbounded. This follows from two things: (i) IIT holds that 'level of consciousness' is measured by Φ, and (ii) Φ has a ratio scale and is unbounded. In his commentary, Aaronson (2014) implicitly agrees that IIT has these implications. He writes that the theory 'predicts that [a] Vandermonde contraption would be *billions of times more* conscious than you are'. This assumes a ratio scale. He also writes that IIT implies that such contraptions 'can be *unboundedly more* conscious than humans are' (my italics) because they can have unboundedly high levels of Φ. And, in his 2014 response, Tononi doesn't dispute these claims.

Now let me turn back to my main question, Q1. What specific dimension do integrated information theorists have in mind by the expression 'level of consciousness' when they propose in their theory that 'level of consciousness is exactly measured by Φ'? Let us go through the candidates listed above one by one.

To begin with, it is quite clear that integrated information theorists do not have in mind either Intensity Level or Complexity Level when they speak of 'level' or 'amount' of consciousness. For instance, Tononi (2014) says that if you stare at blank wall then you can be 'highly conscious'. But the qualitative content of this experience is neither intense nor complex.

But then what do integrated information theorists have in mind by 'level' or 'amount' of consciousness? One clue is provided by the following passage from Tononi and Koch's paper 'Consciousness: Here, There and Everywhere?':

> **P1:** Consciousness is graded… In us [consciousness] becomes richer as we grow from a baby to an adult whose brain has fully matured and becomes more functionally specialized. It can also wax and wane when

> we are highly alert or drowsy, intoxicated by drugs or alcohol, or become demented in old age. (Tononi and Koch, 2015, p. 11)

They then speculate that the waxing and waning of 'level of consciousness' correlates with the waxing and waning of Φ. So here is one thing we know about 'level of consciousness' as it is understood by Tononi and Koch: it varies with age, level of alertness, and level of intoxication.

This is consistent with their having in mind either Determinacy Level or Access Level. For it is plausible that these also wax and wane with level of alertness, intoxication, and so on. For instance, when we become very intoxicated or sedated, the contents of our experiences may become less determinate. In addition, less information from the content of experience is cognitively processed for use by 'consuming systems'. That is, it is not implausible that both Determinacy Level and Access Level go down. Maybe, then, we should interpret their theory as claiming that one of *these* is determined by Φ?

The trouble is that if Tononi and Koch mean either one of these things then IIT would have false implications. As noted above, IIT implies that 'level of consciousness', whatever it is, is unbounded, because Φ is unbounded. Indeed, Aaronson shows that a large network of XOR gates arranged in a simple expander graph can have arbitrarily high levels of Φ. By contrast, in any possible creature, both Determinacy Level and Access Level have an *upper bound*. In principle, there is always a maximum for Determinacy Level: it occurs when a creature has an experience as of a property and there are no more specific, perceptually available ways of having that property. For instance, if you look at a white wall, maybe the qualitative content of your experience could involve a maximally-specific shade of colour. Also, there is always a maximum for Access Level: for instance, if you are fully awake and attentive, you might fully cognitively 'take in' the complete content of experience while you are having the experience. No higher level of access is possible. So if the IIT 'level of consciousness is determined by Φ' is interpreted to be about either of Determinacy Level or Access Level, its implication about unboundedness is false.

A further point adds to the case against interpreting 'level of consciousness' as Access Level. Recall that IIT implies that an expander graph might have an extremely high 'levels of consciousness' (whatever that means), because it has a high level of Φ. Indeed, its level of consciousness could be unboundedly higher than your own as you look at a white wall. If we take this prediction to mean that its

cognitive access to the contents of its experiences is greater than your own, then the prediction is straightforwardly false. For the 2D grid doesn't have a cognitive system at all! Apparently, for integrated information theorists, 'level of consciousness' picks out some dimension having to do with 'phenomenal consciousness', where the theory implies that this is totally separable from 'cognitive access'.

In sum, when integrated information theorists declare that 'level of consciousness' is determined by Φ, charity of interpretation demands that we *not* take them to be referring either to Determinacy Level or to Access Level. But then what are they referring to?

There is only one remaining candidate on our list: Richness of Experiential Repertoire. Perhaps, when integrated information theorists say that 'level of consciousness is determined by Φ', they should be taken to be asserting something along the following lines: *the number of distinct experiences a system can have* is determined by its level of Φ. Call this the *Repertoire interpretation*.

The Repertoire interpretation fits better than the previous interpretations with the implications of IIT. In particular, we saw that IIT implies that 'level of consciousness' is unbounded. It is not unnatural to think that 'Richness of Experiential Repertoire' is unbounded: there is no upper bound to the number and variety of distinct experiences a system could enjoy.

In addition, some remarks by Tononi in his paper 'Consciousness as Integrated Information: A Provisional Manifesto' suggest something in the vicinity of the Repertoire interpretation. Here are the relevant passages:

> **P2:** You are facing a blank screen that is alternately on and off, and you have been instructed to say 'light' when the screen turns on and 'dark' when it turns off… For you, a light screen is different not only from a dark screen, but from a multitude of other images, so when you say 'light', it really means this specific way versus countless other ways, such as a red screen, a green screen, a blue screen, this movie frame, that movie frame, and so on for every movie frame (not to mention for a *sound, smell, thought, or any combination of the above*)… All this added meaning, provided implicitly by how we discriminate pure light from all these alternatives… increases *the level of consciousness*. (Tononi, 2008, pp. 217–18, my italics)

> **P3:** By subtraction, [you] can realize that, if [you] were to lose one neural mechanism after the other, [your] being conscious of 'light' *would degrade* — it would lose its non-coloredness, its non-shapedness, it would even lose its visualness… (*ibid.*, p. 218, my italics)

> **P4:** The IIT claims that consciousness is not an all-or-none property, but is graded: specifically, it increases in proportion to a system's repertoire of discriminable states. (*ibid.*, p. 236)

However, other considerations strongly count against the Repertoire interpretation.

(i) The passages P2 and P3 are extremely hard to understand. But the most natural interpretation is this: when you experience a white screen, your 'level of consciousness' at that time is determined by the number of distinct, alternative experiences you could have at that time, given your neural machinery.

But this is an exceedingly odd notion of 'level of consciousness'. For here is a natural assumption: your experience of the white screen could stay the same, while there is variation in the number of *alternative* experiences you could have. For instance, while having the *very same* experience of the white screen, you could (unknown to you) suddenly totally lose the capacity to have any sensations of smell. After all, people do entirely lose their sense of smell (due to an infection, injury, or stroke) but retain the capacity to have *exactly the same* visual experiences that they had before. Then the suggested notion of 'level of consciousness' undeniably implies that your 'level of consciousness', while experiencing the white screen, would go down, even though the *phenomenal character of your total experience has remained exactly the same*, because of a reduction in the number of *alternative* experiences you can have. That is, your 'level of consciousness' at a time is independent of the phenomenal character of your total conscious state at that time. And this is a very odd conception of 'level of consciousness'.

(ii) Although passages P2–P4 from Tononi's earlier 2008 essay suggest the Repertoire interpretation, other passages in Tononi and Koch's more recent 2015 essay mentioned above seem to rule it out. For instance, consider passage P1 above from that essay. In this passage Tononi and Koch say that drowsy, intoxicated, and demented people have a 'lower level of consciousness' than the rest of us. (They speculate that this is because they have lower levels of Φ, in accordance with IIT.) But a drowsy person could presumably experience all the colours, smells, and so on that the rest of us can experience. After all, her sensory systems work just fine. True, she may not cognitively respond to them as much as us; but just because there is a reduction in cognitive access does not mean there is a reduction in phenomenal consciousness. (Likewise, in at least *some* cases of intoxication or dementia, there is no reduction in the number of basic sensations the

person can have.) Since Tononi and Koch say in this passage that a drowsy person has a 'lower level of consciousness' than a fully awake person, but since she doesn't have a lower repertoire of possible experiences, this passage suggests that they do *not* have the repertoire interpretation of 'level of consciousness' in mind. But then we are left wondering what they mean.

In addition, in the same 2015 essay, Tononi and Koch say that 'an experience of pure darkness and silence is what it is because… it necessarily differs from a large number of alternative experiences I could have' (p. 6). This is similar to what Tononi said in the passages above from his 2008 essay. But, in their 2015 paper, they do *not* call this 'level of consciousness'; in this essay, they call it 'information'. They do use 'level of consciousness' elsewhere in the essay, but they do not explain what they mean. So, in this paper at least, they appear to use 'level of consciousness' to mean something *other than* repertoire of alternative experiences. But, again, we then are left wondering what they mean.

Likewise, elsewhere Tononi (2012, p. 306) says: 'the "richness" of an experience is the number of [experiential] dimensions', while 'the level of consciousness is the value of maximally integrated conceptual information'. This clearly shows that he takes 'level of consciousness' to refer to something *other than* the number of potential experiences.

In sum, when integrated information theorists use 'level of consciousness', it is not clear that they have in mind *any* one of the dimensions listed above. Perhaps, then, they mean some *other* dimension, not yet specified? But it is hard to see what this other dimension could be since the list above seems to exhaust the options. Or perhaps their use of the term is vague or indeterminate: they use the term but they simply don't have in mind anything in particular.

In any case, as long as integrated information theorists do not address the question, we literally do not know what they are claiming. It is as if they are saying 'the bling-value of a system's consciousness is measured by the Φ-value of that system', without giving any indication of what they mean by 'bling-value'.

I am not saying that the theory 'the *amount* or *level* of a system's consciousness is determined by its Φ-value' is *false.* My worry is that

no clear view has yet been put on the table that we can test as true or false.[6]

2. How Might IIT Explain Qualitative Content?

At the start of the previous section, I noted that one question for a theory of the physical basis of consciousness is the *Ignition Question*: *when* is a physical system conscious? This is the question that has received the most attention. For instance, global broadcast theories (Baars, 1988; Dehaene, 2014) address the Ignition Question only. But there is an additional question a complete theory must answer:

[6] It may be worth mentioning yet another, somewhat unusual, interpretation of integrated information theorists' talk of 'the level of consciousness'. I will call it the *inscrutable interpretation*. Let me introduce it with an analogy. Suppose that you belong to a species that is like us except that the species has no vision whatever. However, suppose that by some miracle you occasionally have a ganzfeld experience of a certain specific shade of red. You never experience any other colours. Then you might not appreciate that there is a general dimension *colour* which encompasses many other colours. A neuroscientist might somehow infer that there is such a dimension, but you would have no direct grip on it. You would be 'cognitively closed' with respect to this dimension. On the inscrutable interpretation, we are in a similar position with respect to the relevant dimension of consciousness. That is to say, since integrated information theorists are wedded to their theory that consciousness is attached to Φ, and since different systems than us (e.g. 2D gates) have different levels of Φ, they infer that they must enjoy different 'levels' of consciousness than we do. But, according to the inscrutable interpretation, because we are at all times 'stuck' with our own, single level of consciousness, we have no *direct grasp* of this dimension and the other possible values along this dimension. So, 'level of consciousness', the variable magnitude allegedly measured by Φ, should not be taken to refer to one of the dimensions discussed above which we do grasp; instead, it refers to another, hitherto undiscovered, dimension that we do *not* directly grasp. Our only reason to believe in this 'other dimension' is based on the speculative integrated information theory itself. Could this be the correct interpretation of IIT as defended by Tononi and Koch? To begin with, it has an odd consequence: IIT is, in the first instance, a theory of something we don't have a direct grasp on! For the idea is that, although we have a grasp on our own consciousness, we have no direct grasp on the relevant magnitude *level of consciousness* — just as in the example above, although we have a grasp on the relevant shade of red, we have no direct grasp on the general dimension *colour*. In any case, the inscrutable interpretation is directly contradicted by what Tononi and Koch say. For, according to the inscrutable interpretation, we are *at all times* stuck with exactly the same level of consciousness — this is supposed to be why we have no direct grip on this dimension. As against this, in passage P1 above, Koch and Tononi say that our own level of consciousness waxes and wanes with our age, with our level of intoxication, and so on. (Indeed, this is crucial to the possibility of empirically confirming the theory in the first place, because in that case the theory would be confirmed if Φ waxes and wanes in the same way.) So, from such waxing and waning in our own case, we should have a direct grasp of at least some different levels of consciousness. What Tononi and Koch say here, then, is directly inconsistent with the inscrutable interpretation.

> *The Quality Question*: *What* specific conscious experiences does a physical system have? More exactly, what is the *complete set of psychophysical principles* which, together with the physical facts concerning any arbitrary physical system, entail the precise qualitative contents of all of its conscious experiences?

To illustrate, suppose that you are a super-scientist, you know all the physical facts about some unfamiliar sentient organism, Karl. If you also somehow knew the complete set of psychophysical principles, you could deduce the qualitative contents of all of Karl's experiences.

For reductionists about consciousness, such principles will be underwritten by identities between states of consciousness and complex physical states (e.g. neural states). For those who consider consciousness to be an emergent phenomenon, the principles will be basic 'laws' relating physical states (e.g. neural states) with distinct, emergent states of consciousness. The holy grail would be psychophysical laws with the simplicity and elegance of the basic laws of physics.

Many philosophers (e.g. Dretske, 1995; Tye, 2000) favour an *externalist* answer to the Quality Question. This view has two parts. First, there is a raft of basic identities between 'qualia' and physical properties in the world: for instance, shape qualia are just physical shapes, colour qualia are certain reflectance-types, smell qualia are chemical types, and so on. Second, we experience them by having neural states that have the biological function of indicating their occurrence.

In my view, however, such a view faces huge empirical and *a priori* problems. I believe that standard neuroscience provides strong empirical support for an *internalist* answer to the Quality Question.[7] In particular, the qualitative contents of experience are completely determined by our internal neural states. Let me provide some examples.

[7] For some *a priori* problems with Dretske-Tye externalism, see Pautz (2017). For the empirical problems, see Pautz (2014). The basic problem is that there is no systematic isomorphism or mapping between qualia and external physical properties (e.g. very similar chemical types often give us experiences of very different smell qualia). By contrast, as I am about to point out, there is a better correlation between qualitative content and neural patterns. Papineau (2003) expresses scepticism about whether the debate between externalism and internalism about qualitative content can be settled empirically. (Papineau, 2016, instead offers a non-empirical, philosophical argument against Dretske-Tye externalism and for an internalist view.) As against Papineau's scepticism, such empirical findings clearly support internalism over externalism.

Afterwards we will look at integrated information theorists' very different answer to the Quality Question.

Pain intensity. Using noxious temperatures and measuring neural activity with fMRI, Coghill *et al.* (1999, p. 1936) found that 'many cortical areas exhibit significant, graded changes in activation *linearly related to* pain intensity'. Kenshalo *et al.* (2000) found in a single-unit study that the relationship between temperature and the firing rates of wide dynamic range (WDR) neurons in monkey S1 very closely resembles the psychophysically-derived relationship between temperature and pain intensity in humans. And Timmerman *et al.* (2001) found using magnetoencephalography that S1 neural activity perfectly matched subjects' pain ratings in response to nociceptive laser stimuli delivered to the hand (see Figure 1a).

Smell and taste. Howard and co-workers (2009) found that 'spatially distributed ensemble activity in human posterior piriform cortex (PPC) coincides with perceptual ratings of odor quality, such that odorants with more (or less) similar fMRI patterns were perceived as more (or less) alike'. For instance, they found that your PPC neural representation of *R*-limonene resembles your PPC neural representation of citral more than your PPC neural representation of *R*-carvone, in perfect agreement with the character of your smell experiences. Similar results have been found for the experience of taste (Crouzet, Busch and Ohla, 2015).

Audition. Relkin and Doucet (1997, p. 2738) write that 'the perceived loudness of a pure tone appears to be linked both to the number of spikes fired by single neurons and to spatial spread of excitation in the auditory nerve'. Langers *et al.* (2007) used fMRI to look at neural activity further downstream in the auditory cortex. They found that 'cortical activity is more closely related to the perceptual loudness level of sound than to its [external, physical] intensity level' (p. 714) and indeed report 'a type of non-linearity… comparable to that reported in psychophysical studies on loudness perception that employ subjective loudness scaling' (*ibid.*, p. 716). On the basis of this study and others, Röhl, Kollmeier and Uppenkamp (2011, p. 1494) conclude that 'the most simple interpretation would be that AC [auditory cortex] is fed by… the auditory brainstem according to the sound pressure level and the bandwidth of the stimuli, and an additional component is added which is linearly related to the perceived loudness'.

Colour. In a recent study, Bohon *et al.* (2016) recorded the activity of neurons in V4 (see also Brouwer and Heeger, 2009). They then

used multidimensional scaling to analyse their colour-tuning. Here is how they summarize their results:

> The arrangement of the [neural responses] clearly reflects color space: points of the same hue irrespective of luminance level are plotted next to each other, and the progression of the points forms a circle that proceeds according the color wheel. Behavioral judgments of the similarity between colors closely match the similarities between the neural responses to these colors by the glob neural population. (p. 18)

See Figure 1b for an illustration. There is also the fact that we experience unitary colours (red, green, yellow, blue) and binary colours (orange, purple, etc.). The textbook explanation is that we have an r-g channel and a y-b channel. The experience of unitary colours corresponds to equilibria states of the channel and the experience of binary colours correspond to a departure from equilibria in both channels. This explanation is controversial (Webster, 2018), but it does continue to have adherents (Schmidt, Neitz and Neitz, 2014; Danilova and Mollon, 2012).

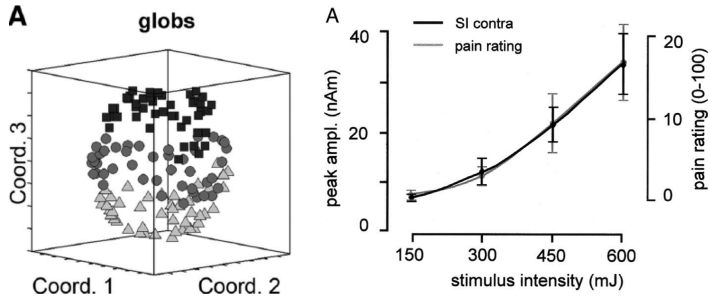

Figure 1. *A.* Multidimensional scaling and representational similarity analyses showed that neural pattern in both glob and interglob populations were correlated with the organization of CIELUV space, but glob cells showed a stronger correlation. From Bohon *et al.* (2016, Creative Commons). *B.* Amplitudes of SI activity (black line, left scale) match precisely the subjects' pain ratings (grey line, right scale). Reprinted from Timmerman *et al.* (2001) with permission.

Now these studies fall short of providing a complete set of psychophysical principles. But they do offer some rough principles. For instance, the intensity of qualitative content is determined by firing rates. The similarity of experiences is determined by similarity of distributed neural patterns, as measured by multidimensional scaling. Whether you experience a unitary or binary colour is perhaps determined by relative activity across the r-g and y-b opponent channels.

The hope is that there is a complete set of precise psychophysical principles in the vicinity. Maybe, for instance, there is one psychophysical principle for colour, another for smell, yet another for pain intensity, and so on. There might be a number of systematic mappings from neural parameters onto qualitative parameters. Each of them might have something like the following form:

> [L] If a system has neural state *P*, then the system has an experience with qualitative content *f*(*P*).

Here *f* is a systematic function or mapping from specific neural states onto specific qualitative contents. As Stanislas Dehaene writes, 'the [neural] code contains a full record of the subject's experience' and 'if we could read this code we should gain full access to a person's inner world' (2014, pp. 143–5). Let us call this the *standard neuroscience* answer to the Quality Question.[8]

The answer to the Quality Question offered by integrated information theorists is quite different. While standard internalism holds that qualitative content is systematically determined by *distributed, spatio-temporal patterns of neuronal firing*, integrated information theorists hold that it is determined by something that I will call 'T-shape' (for *Tononi shape*), leaving it neutral whether T-shape is the basis of qualitative content. This is a novel notion that has not figured in standard neuroscience. Very roughly, a distributed pattern of neural activity can help determine a 'T-shape'. This is a literal shape in a multi-dimensional 'cause–effect space'. The T-shape is not fixed by the distributed pattern of neural activity alone. Instead, it is determined *holistically*:

> Rather than trying to understand the meaning of the activity of some elements (neurons) in isolation, or even of distributed patterns of activity, the IIT claims that meaning is only generated in terms of shapes in Q, that is, *in terms of the set of informational relationships generated by a complex…* Moreover, informational relationships, and thus the [Q-shape] are specified both by the elements that are firing and by those that are not. (Balduzzi and Tononi, 2009, p. 12)

[8] There are many neglected challenges to the idea that there are systematic psychophysical mapping principles entailing the qualitative contents of any creature given the physical facts about that creature. See Adams (1987, pp. 256–7), Chalmers (2012, pp. 279, 341), MacLeod (2010), O'Regan (2011, p. 99), Prinz (2012, pp. 126–33), and Teller and Pugh (1984). Notwithstanding these challenges, standard neuroscience has made real progress when it comes to the Quality Question.

In other words, a T-shape 'embodies the entire set of informational relationships generated by interactions in the system' (*ibid.*, p. 1). It follows that 'it does not make sense to ask about the [T-shape] generated by a mechanism in isolation, or by a state (firing pattern) in isolation'. In fact, 'two different systems having identical activity patterns may generate different [T-shapes]', because of holistic differences in informational relationships (*ibid.*).

For a system with n binary elements and 2^n possible states, T-shape is a shape in a 2^n dimensional space (*ibid.*, p. 6). For a human brain, which has endlessly many possible states, T-shape will be a shape in a space with a mind-boggling number of dimensions. So integrated information theorists concede that 'in practice it is not possible to determine [T-shape] precisely for a realistic system' (*ibid.*, p. 9).

Nevertheless, integrated information theorists make a bold speculation. They posit a psychophysical principle linking these hypothesized 'T-shapes' (which are physical, albeit abstract, properties) with qualitative content:

> *Similarity-Congruence:* If the total state of a system determines T-shapes that stand in a certain similarity-order (e.g. T_1 is more than T_2 than T_3), then the system experiences qualitative contents that stand in the same similarity-order.

For instance they write, 'similarity between experiences reduces to [or is determined by] similarities between shapes' (*ibid.*, p. 23). They have not said much more than this on the issue of how the physical facts determine the specific qualitative contents of our experiences.[9]

In sum, while integrated information theorists hold that Φ determines 'level of consciousness', they hold that 'T-shape' determines the specific qualitative content of experience.

Previously, I just assumed that Φ is a well-defined physical quantity. Then I raised some basic questions about the further claim that it determines 'level of consciousness'. Likewise, for the sake of

[9] Here is another quotation: 'Different experiences — every different scene in a movie or in a dream — correspond to different [T-]shapes, with some shapes being measurably closer (red and blue) and some more distant within the space (a black screen and a city scene)… Indeed, there is much scope for future research to begin mapping psychophysics, for example, the circular nature of colour space, onto the geometry of shapes in cause–effect space — except that a shape in cause–effect space, unlike the shape of an object in 3D space, is the shape within, the shape of experience itself… It is the voice in the head, the light inside the skull' (Tononi and Koch, 2015, p. 11).

discussion, I will now assume that T-shapes are well-defined, abstract physical properties of a physical system. This leaves open the further issue of whether T-shapes determine the qualitative contents of experience. I have a number of basic questions about this idea.

To begin with, recall that what is needed is a *complete set of psychophysical principles*: a set of principles which, together with the physical facts concerning any physical system, entail the precise qualitative contents of all of its conscious experiences. But proponents of IIT have only proposed a single psychophysical principle: Similarity-Congruence. For several reasons, this falls well short of a complete set of psychophysical principles. And it is not clear how the single notion of 'T-shape' could be the basis of a complete psychophysical theory. Let us take these points in turn.

To see how Similarity-Congruence falls short, let us return to the hypothetical case mentioned above: you are a super-scientist, you know all the physical facts about some unfamiliar sentient organism, Karl. You want to determine exactly what experiences Karl has. That is, you are engaged in the task of 'radical phenomenal interpretation'. Suppose that Karl is presented with three objects consecutively and, given the physical facts about his brain, you deduce that it is associated with three T-shapes such that T_1 is more like T_2 than T_3. Then, given Similarity-Congruence, you can deduce *that Karl has some trio of experiences, E_1, E_2, and E_3, such that E_1 is more like E_2 than E_3*. But, as a simple point of logic, Similarity-Congruence is not logically strong enough to tell us precisely *what* those experiences are. For instance, it doesn't tell us whether they are colour experiences of similar shades of *red*, or whether they are colour experiences of similar shades of *green*. In fact, it doesn't tell us whether they are experiences of *colour* or experiences of (say) *smell*. That is, it doesn't entail the *specific, determinate* qualitative contents of those experiences. This is not to say that Similarity-Congruence is false; it is just to say that it is not the full story.

Here is a second respect in which Similarity-Congruence falls short. Take colour experiences. The qualitative contents of colour experiences vary along several dimensions: *hue* (red vs. green and yellow vs. blue), *saturation*, and *brightness*. So it is natural to expect that there are corresponding physical parameters determined by neural activity and that there are systematic psychophysical laws going from these parameters to the qualitative dimensions. As we saw above, the models of standard neuroscience roughly provide such laws (e.g. the

hypothesized r-g and y-b channels determine hue). But Similarity-Congruence provides nothing of the sort.

Finally, the qualitative contents of our experience do not just stand in similarity relationships; they also vary in *intensity* and indeed some exhibit a ratio scale. For instance, suppose again you are a super-scientist observing Karl. Karl experiences two tones where the second tone is 10 dB more intense than the first one. For humans, a general rule of thumb is that a 10 dB increase in intensity causes a doubling in perceived loudness. Suppose that this holds for Karl as well. Then here is a phenomenal fact about his experiences: *the loudness-level that Karl experiences roughly doubles*.[10] A complete set of psycho-physical principles should entail this fact, given a complete physical description of Karl's neural response to the 10 dB increase. But, as a simple point of logic, nothing like this can be derived from the physical facts about Karl and Similarity-Congruence alone.

So here is my first question:

> **Q4:** Since Similarity-Congruence falls short, how might psycho-physical principles based on the notion of T-shape be formulated, in order to obtain a complete, systematic theory? Given a physical description of Karl, such principles would entail exactly what experiences he has (colour experiences, olfactory experiences, etc.), their intensities, and so on.

How might proponents of the T-shape theory answer this question? Balduzzi and Tononi make a speculation that may seem to help:

> According to the IIT, phenomenological differences [between the experience of colour, smell, and so on] correspond to different basic sub-shapes in Q, such as grid-like structures and pyramid-like structures. (Balduzzi and Tononi, 2009, p. 20)

> In the IIT framework, colors correspond to different sub-shapes of the same kind (say pyramids pointing in different directions) and sounds to very different sub-shapes in Q. (*ibid.*, p. 22)

The idea seems to be this. We have seen that integrated information theorists speculate that the brain can be associated with T-shapes in a space with a mind-boggling number of dimensions. Now they are adding that the T-shapes fall into categories. To illustrate, they say

[10] Here I am assuming that perceived loudness has a genuine ratio scale so that it makes sense to say that perceived loudness doubles with a 10 dB increase. One piece of evidence for this is that subjects' estimations obey additivity (Gescheider, 1997, p. 265).

that different pyramids may be the physical basis of different colour experiences; T-shapes of a different kind may be the physical basis of auditory experiences; and so on. Of course, 'pyramids' is just a somewhat silly example used for illustrative purposes. The idea is that there are *some* different categories of high-dimensional T-shapes that underlie the different categories of experiences, but we do not know what they are.

Of course, this is another piece of unbridled speculation. But assuming for the sake of discussion that it is true, we can ask how it could be the basis of a simple set of complete psychophysical principles.

Maybe the idea is that the final complete psychophysical theory will replace Similarity-Congruence with a swarm of more specific principles invoking the different categories of T-shapes. But what are these more specific principles?

One idea is that they are more specific principles of the same form as Similarity-Congruence: 'similar pyramid shapes are the basis of the experience of similar phenomenal colours' and 'similar grid-shapes are the basis of the experience of similar smells', and so on. But this proposal still falls short of completeness in the ways catalogued above. These principles are still too unspecific. For instance, if you were a super-scientist and knew that Karl's brain state on two occasions determined similar pyramid shapes, then from such principles you would only be able to deduce that he experiences *some or other* similar phenomenal colours on those occasions, but you would not be able to deduce *exactly what* these phenomenal colours are.

To guarantee completeness, integrated information theorists might suggest a list of even more specific fundamental principles that cover all the possible cases:

> Pyramid shape S_1 is the physical basis of experiencing red_1
> Pyramid shape S_2 is the physical basis of experiencing red_2
> …
> Pyramid shape S_{101} is the physical basis of experiencing $green_1$
> …
> Grid-like shape S_1 is the physical basis of experiencing $smell_1$
> Grid-like shape S_2 is the physical basis o experiencing $smell_2$
> Etc.

This would indeed guarantee completeness. If you were a super-scientist and you were armed with a list of such principles covering all possible cases then you could deduce the specific qualitive contents of

all of Karl's experiences — or indeed the qualitative contents of any possible sentient being.

But it is hard to believe that there is a different *fundamental* principle for each possible experience. For one thing, this view is complicated and unlovely. A much simpler psychophysical theory would posit a handful of general, systematic psychophysical principles, from which such more specific principles could be derived as special cases. Compare how Newton's law of gravitation implies many specific mass–gravity connections as special cases. For another thing, there is the point already made above. The qualitative contents of our experiences vary along certain dimensions (hue, saturation, brightness, intensity, etc.). So it is natural to expect that there are corresponding physical parameters determined by neural activity and that there are systematic psychophysical laws going from these parameters to the qualitative dimensions. Thus, a list of associations between specific physical states and specific experience, such as the list above, cannot be the full story. There is reason to think that it is possible to specify more systematic mappings from physical states onto experiences.

In sum, what proponents of the T-shape theory of qualitative content need is to at least sketch how a psychophysical theory based on T-shape could be at once complete *and systematic*. Given the physical facts about Karl, such a complete set of systematic psychophysical principles would have to be powerful enough to entail, for instance, that *the loudness-level that Karl experiences roughly doubles with a 10 db increase.* Standard neuroscience *can* sketch such principles. For instance, as noted above, Relkin and Doucet (1997, p. 2738) suggest that 'the perceived loudness of a pure tone appears to be linked both to the number of spikes fired by single neurons and to spatial spread of excitation'. My question Q4 above is: can proponents of the T-shape theory sketch such principles? For instance, what is it about these high-dimensional T-shapes that could systematically determine the *loudness* of experienced sound?[11]

[11] In discussion, Kelvin McQueen has suggested that perhaps integrated information theorists could say that the T-shapes underlying the experience of sounds have 'extents' and the 'extent' of the T-shape determines the perceived loudness of sound. Thus, if the 'extent' doubles, the loudness of the perceived sound doubles. This would be an alternative to the standard neuroscience explanation in terms of firing rates and extent of excitation. But I am not sure what is meant by 'the extent' of a T-shape. In addition,

So much for Q4. My next question is this:

> **Q5:** According to the T-shape theory, what is the physical basis of 'binding'?

For instance, suppose that Karl hallucinates a blue sphere above a green triangle. Presumably, on the T-shape theory, Karl's brain determines four T-shapes, which are the physical bases of experiencing blue, experiencing green, experiencing roundness, and experiencing triangularity. But this does not explain why Karl experiences these features as bound together in the way that he does. Again, what is needed is a sketch of a complete set of systematic principles that entails *all* the experiential facts about Karl given the physical facts about him, including patterns of binding.[12]

Here is my final question:

> **Q6:** Is there any empirical evidence that supports the T-shape theory of qualitative content over standard neuroscience, which does not appeal T-shape?

As far as I can tell, at present the answer to this question is no. At this stage, the idea that there could be a complete and systematic answer to the Quality Question based on the notion of T-shape is totally speculative. As noted above, for a human brain, which has endlessly many possible states, T-shape will be a shape in a space with a mind-boggling number of dimensions. So integrated information theorists concede that 'in practice it is not possible to determine [T-shape] precisely for a realistic system' (Balduzzi and Tononi, 2009, p. 9). So it may be that the T-shape theory is untestable. True, Tononi and Koch rather hopefully say that 'there is much scope for *future* research to begin mapping psychophysics, for example, the circular nature of colour space, onto the geometry of shapes in cause–effect space' (2015, p. 11, my italics). But, at present, this remains totally untested. So, for instance, it is a prediction of the T-shape theory that, since the experience of blue is more like the experience of purple than the

unlike the standard neuroscience explanation, this explanation has not been supported empirically.

12 Tononi *et al.* (2016, p. 457) do offer one remark on this issue: 'According to IIT, this dynamic binding of phenomenal attributes occurs if, and only if, in cause–effect space the corresponding concept purviews are related, meaning that they refer to an overlapping set of PSC elements and jointly constrain their past or future states.' This will perhaps be elaborated in future work.

experience of green, the corresponding T-shapes stand in an isomorphic resemblance-order. But there is no evidence to back this up. By contrast, as noted above, standard neuroscience has made real progress on the Quality Question, without invoking 'T-shapes' in an unspecified high-dimensional space.[13]

It seems to me that one apparent empirical prediction of the T-shape theory is already known to be false. Balduzzi and Tononi (2009, p. 1) write that a T-shape 'embodies the *entire* set of informational relationships generated by interactions in the system' (my italics). Accordingly, Tsuchiya (2007, p. 7) writes that it is a prediction of the T-shape theory that 'the quality of experience… can be determined only by the interactions with other neural interactions in a holistic manner' so that the '*visualness* of visual experience is determined not only by the way visual neurons interact with other visual neurons, but it also depends on how the visual neurons interact with *auditory neurons* and *other neurons* within the complex' (*ibid.*, my italics). In other words, the T-shape theory predicts that 'vision cannot feel like vision unless it is related with other senses' (*ibid.*, Figure 2, p. 6). Now, is this prediction correct? True, there are *some* cross-modal interactions. For instance, because of cross-modal binding, if you became totally deaf, the phenomenal character of *some* your visual experiences might be different. For instance, visually perceived dogs would no longer appear to make barking sounds. But the quotations suggest something much stronger: a radical form of experiential holism. For instance, taken literally, they imply that when a person entirely loses the capacity to experience smell (say, due to a stroke) then *all* of her visual experiences of the world change at least *somewhat* in their phenomenal character. Tononi (2008, pp. 217–18) says this explicitly regarding his 'subtraction' thought experiment. However, this prediction is false. Of course, Balduzzi and Tononi (2009) and Tsuchiya (2007) may disavow my literal interpretation of their remarks. In that

[13] Tsuchiya (2017, Section 4.1, p. 7) claims to have found empirical evidence for the T-shape theory of qualitative content. But, at present, there is not sufficient empirical evidence to support the T-shape theory of qualitative content over the quite well-confirmed explanations offered by standard neuroscience that were discussed above, which do not appeal to T-shape. As discussed above, multiple studies and decades of research support these explanations of the experience of pain intensity, loudness, smell and taste, and colour.

case, they need to clarify what they mean and state more clearly the empirical predictions of the T-shape theory.[14]

3. Summary

In the first place, IIT is a theory of the 'level' or 'amount' of consciousness in a system. But it is totally unclear what this means because consciousness varies along many dimensions. As long as integrated information theorists do not address the question, we literally do not know they are claiming. It is as if they are saying 'the bling-value of a system's consciousness is measured by the Φ-value of that system', without giving any indication of what is meant by 'bling-value'. Until they say more, no clear, testable theory has been put on the table.

Integrated information theorists also offer a theory of the specific qualitative contents of experiences in terms of 'T-shapes'. This theory faces several basic questions. It is hard to see how there could be a complete and systematic psychophysical theory based on the notion of T-shape. Moreover, the theory is at this point totally speculative and empirically untested. If anything, its prediction of radical holism is already known to be false. By contrast, standard neuroscience has made real progress on explaining the qualitative content of experience, without invoking 'T-shapes'.

Acknowledgments

This essay develops some questions I posed in Pautz (2015) in the wake of a workshop on integrated information theory at New York University. My thanks to Tim Bayne, Ned Block, Kelvin McQueen, and two anonymous referees for very helpful suggestions that led to improvements.

References

Adams, R. (1987) Flavors, colors, and god, in *The Virtue of Faith and Other Essays in Philosophical Theology*, Oxford: Oxford University Press.

[14] Balduzzi and Tononi (2009, p. 15) point out another fascinating prediction of the T-theory: that a 'frozen' brain undergoing *no neural activity* at all still would have certain alien experiences that we cannot imagine. This is harder to test! But, since all the evidence suggests that neural firing is required for sensation, one cannot be blamed for finding this prediction highly dubious at best.

Aaronson, S. (2014) Why I am not an integrated information theorist (or, the unconscious expander), *Shtetl-Optimized: The Blog of Scott Aaronson*, [Online], http://www.scottaaronson.com/blog/?p=1799 [23 Sept 2018].
Baars, B.J. (1988) *A Cognitive Theory of Consciousness*, Cambridge: Cambridge University Press.
Balduzzi, D. & Tononi, G. (2009) Qualia: The geometry of integrated information, *PLoS Computational Biology*, **5** (8), pp. 1–24.
Bayne, T., Hohwy, J. & Owen, A. (2016) Are there levels of consciousness?, *Trends in Cognitive Sciences*, **20** (6), pp. 405–413.
Block, N. (2007) Consciousness, accessibility, and the mesh between psychology and neuroscience, *Behavioral and Brain Sciences*, **30**, pp. 481–499.
Block, N. (2019) Fading qualia: A response to Michael Tye, in Pautz, A. & Stoljar, D. (eds.) *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*, Cambridge, MA: MIT Press.
Bohon, K.S., Hermann, K.L., Hansen, T. & Conway, B.R. (2016) Representation of perceptual color space in macaque posterior inferior temporal cortex (the V4 complex), *eNeuro*, **3** (4), ENEURO-0039.
Brouwer, G. & Heeger, D. (2009) Decoding and reconstructing color from responses in human visual cortex, *Journal of Neuroscience*, **29**, pp. 13992–14003.
Casali, A., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K., *et al.* (2013) A theoretically based index of consciousness independent of sensory processing and behavior, *Science Translational Medicine*, **5** (198).
Chalmers, D. (1995) *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.
Chalmers, D. (1998/2010) On the search for the neural correlate of consciousness, in *The Character of Consciousness*, New York: Oxford University Press.
Chalmers, D. (2006) Perception and the fall from Eden, in Gendler, T. & Hawthorne, J. (eds.) *Perceptual Experience*, Oxford: Oxford University Press.
Chalmers, D. (2012) *Constructing the World*, New York: Oxford University Press.
Chalmers, D. (forthcoming) Extended cognition and extended consciousness, in Colombo, M., Irvine, E. & Stapleton, M. (eds.) *Andy Clark and His Critics*, Oxford: Wiley-Blackwell.
Coghill, R., Sang, C., Maisog, J. & Iadarola, M. (1999) Pain intensity processing within the human brain, *Journal of Neurophysiology*, **82**, pp. 1934–1943.
Crouzet, S.M., Busch, N.A. & Ohla, K. (2015) Taste quality decoding parallels taste sensations, *Current Biology*, **25**, pp. 1–7.
Danilova, M. & Mollon, J. (2012) Cardinal axes are not independent in color discrimination, *Journal of the Optical Society of America*, **29**, pp. 157–164.
Dehaene, S. (2014) *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*, New York: Viking.
Dretske, F. (1995) *Naturalizing the Mind*, Cambridge, MA: MIT Press.
Fodor, J. (2007) Headaches have themselves, *London Review of Books*, [Online], https://www.lrb.co.uk/v29/n10/jerry-fodor/headaches-have-themselves, [24 Sept 2018].
Gescheider, G. (1997) *Psychophysics: The Fundamentals*, Mahwah, NJ: Lawrence Erlbaum Associates.
Howard, J.D., Plailly, J., Grueschow, M., Haynes, J.D. & Gottfried, J.A. (2009) Odor quality coding and categorization in human posterior piriform cortex, *Nature Neuroscience*, **12**, pp. 932–939.

Kenshalo, D.R., Iwata, K., Sholas, M. & Thomas, D.A. (2000) Response properties and organization of nociceptive neurons in area 1 of monkey primary somatosensory cortex, *Journal of Neurophysiology*, **84**, pp. 719–729.

Koch, C. (2009) A 'complex' theory of consciousness, *Scientific American*, [Online], https://www.scientificamerican.com/article/a-theory-of-consciousness/ [24 Sept 2018].

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

Lamme, V. (2006) Towards a true neural stance on consciousness, *Trends in Cognitive Sciences*, **10** (11), pp. 494–501.

Lamme, V. (2010) How neuroscience will change our view on consciousness, *Cognitive Neuroscience*, **1** (3), pp. 204–220.

Langers, D., van Dijk, P., Schoenmaker, E. & Backes, W. (2007) fMRI activation in relation to sound intensity and loudness, *NeuroImage*, **35**, pp. 709–718.

MacLeod, D. (2010) Into the neural maze, in Cohen, J. & Matthen, M. (eds.) *Color Ontology and Color Science*, Cambridge, MA: MIT Press.

Oizumi, M., Albantakis, L. & Tononi, G. (2014) From phenomenology to the mechanisms of consciousness: Integrated information theory 3.0, *Computational Biology*, **10**, pp. 1–25.

O'Regan, J.K. (2011) *Why Red Doesn't Sound like a Bell: Understanding the Feel of Consciousness*, New York: Oxford University Press.

Pautz, A. (2014) The real trouble for phenomenal externalists: New empirical evidence for a brain-based theory of consciousness, in Brown, R. (ed.) *Consciousness Inside and Out: Phenomenology, Neuroscience, and the Nature of Experience*, Dordrecht: Springer.

Pautz, A. (2015) What is integrated information theory a theory of?, *Scientific American*, [Online], https://blogs.scientificamerican.com/cross-check/consciousness-and-crazyism-responses-to-critique-of-integrated-information-theory/ [23 Sept 2018].

Pautz, A. (2017) The significance argument for the irreducibility of consciousness, *Philosophical Perspectives*, **31** (1), pp. 349–407.

Papineau, D. (2003) Could there be a science of consciousness?, *Philosophical Issues*, **13** (1), pp. 205–220.

Papineau, D. (2016) Against representationalism (about experience), *International Journal of Philosophical Studies*, **24**, pp. 324–347.

Prinz, J. (2012) *The Conscious Brain*, New York: Oxford University Press.

Relkin, E.M. & Doucet, J.R. (1997) Is loudness simply proportional to the auditory nerve spike count?, *Journal of Acoustical Society of America*, **101**, pp. 2735–2740.

Röhl, M., Kollmeier, B. & Uppenkamp, S. (2011) Spectral loudness summation takes place in the primary auditory cortex, *Human Brain Mapping*, **32**, pp. 1483–1496.

Rosen, G. (2010) Metaphysical dependence: Grounding and reduction, in Hale, B. & Hoffman, A. (eds.) *Modality: Metaphysics, Logic and Epistemology*, Oxford: Oxford University Press.

Schmidt, B., Neitz, M. & Neitz, J. (2014) Neurobiological hypothesis of color appearance and hue perception, *Journal of the Optical Society of America*, **31**, pp. 195–207.

Sitt, J.D., King, J.R., Naccache, L. & Dehaene, S. (2013) Ripples of consciousness, *Trends in Neuroscience*, **17**, pp. 552–554.

Teller, D.Y. & Pugh, E.N. (1984) Linking propositions in color vision, in Mollon, J.D. & Sharpe, L.T. (eds.) *Color Vision: Physiology and Psychophysics*, London: Academic Press.

Timmermann, L., Ploner, M., Haucke, K., Schmitz, F., Baltissen, R. & Schnitzler, A. (2001) Differential coding of pain intensity in the human primary and secondary somatosensory cortex, *Journal of Neurophysiology*, **86**, pp. 1499–1503.

Tononi, G. (2008) Consciousness as integrated information: A provisional manifesto, *The Biological Bulletin*, **215** (3), pp. 216–242.

Tononi, G. (2012) Integrated information theory of consciousness: An updated account, *Archives Italiennes de Biologie*, **150** (4), pp. 293–329.

Tononi, G. (2014) Why Scott should stare at a blank wall and reconsider (or, the conscious grid), *Integrated Information Theory*, [Online], http://integratedinformationtheory.org/download/conscious_grid.pdf [24 Sept 2018].

Tononi, G. & Koch, K. (2015) Consciousness: Here, there and everywhere?, *Philosophical Transactions of the Royal Society B*, **370**, pp. 1–18.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016) Integrated information theory: From consciousness to its physical substrate, *Nature Reviews Neuroscience*, **17** (7), pp. 450–461.

Tsuchiya, N. (2017) 'What is it like to be a bat?' — a pathway to the answer from the integrated information theory, *Philosophy Compass*, e12407.

Tye, M. (2000) *Consciousness, Color, and Content*, Cambridge, MA: MIT Press.

Webster, M. (2018) Color vision, in Wixted, J.T. (ed.) *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, Hoboken, NJ: John Wiley & Sons.

Zimmer, C. (2010) Sizing up consciousness by its bits, *New York Times*, [Online], https://www.nytimes.com/2010/09/21/science/21consciousness.html [25 Sept 2018].