Neuroscience of Consciousness, 2018, 4(1): niy007

doi: 10.1093/nc/niy007
Opinion Paper

On the axiomatic foundations of the integrated information theory of consciousness

Tim Bayne*

Department of Philosophy, School of Philosophical, Historical and International Studies, 20 Chancellors Walk, Monash University VIC 3800, Australia

*Correspondence address. Department of Philosophy, School of Philosophical, Historical and International Studies, 20 Chancellors Walk, Monash University VIC 3800, Australia. Tel: 61-03-99020185; E-mail: tim.bayne@gmail.com

Abstract

The integrated information theory (IIT) is one of the most influential scientific theories of consciousness. It functions as a guiding framework for a great deal of research into the neural basis of consciousness and for attempts to develop a consciousness meter. In light of these developments, it is important to examine whether its foundations are secure. This article does just that by examining the axiomatic method that the architects of IIT appeal to. I begin by asking what exactly the axiomatic method involves, arguing that it is open to multiple interpretations. I then examine the five axioms of IIT, asking: what each axiom means, whether it is indeed axiomatic and whether it could constrain a theory of consciousness. I argue that none of the five alleged axioms is able to play the role that is required of it, either because it fails to qualify as axiomatic or because it fails to impose a substantive constraint on a theory of consciousness. The article concludes by briefly sketching an alternative methodology for the science of consciousness: the natural kind approach.

Key words: IIT; contents of consciousness; phenomenological axioms; unity of consciousness; consciousness meter; natural kind approach

Introduction

How should the study of consciousness proceed? Is it possible to determine the distribution of consciousness and discover which neonates, brain-damaged individuals, non-human animals and artificially intelligent agents are conscious? Can we explain why consciousness is absent in some conditions and present in others, or why some kinds of brain activity are associated with consciousness whereas others are not?

This article considers an approach to these questions that I call the *axiomatic approach*. The axiomatic approach lies at the foundations of the integrated information theory (IIT), one of the most influential theories of consciousness (e.g. Oizumi *et al.* 2014; Tononi and Koch 2015; Tononi *et al.* 2016). IIT functions as a guiding framework for significant amounts of research into the neural basis of consciousness and for attempts to develop a consciousness meter (Casali *et al.* 2013; Casarotto *et al.* 2016). Thus, it is important to examine whether its foundations are secure. This article does precisely that.

On the Aims and Ambitions of IIT

In order to properly evaluate IIT, one must understand its aims and ambitions. First, IIT is a theory of subjective experience. Unlike certain theories of consciousness (such as the global workspace theory), which can be read as attempting to account only for the functional dimensions of consciousness, IIT is explicitly presented as a theory of phenomenal consciousness. It aims to account for 'what it is like' to 'perceive a scene, to endure pain, [and] to entertain a thought' (Tononi *et al.* 2016, 450).

Second, IIT is a *reductive* theory of consciousness. The reductive nature of IIT has both ontological and epistemological aspects. Ontologically, IIT purports to provide an account of the fundamental nature of consciousness, claiming that

Received: 24 January 2018; **Revised:** 21 May 2018. **Accepted:** 23 May 2018

consciousness *is* integrated information. Epistemologically, the architects of IIT claim that it 'addresses the hard problem of consciousness in a new way' (Tononi *et al.* 2016, 450).

Third, IIT aspires to be a *comprehensive* theory of consciousness. IIT isn't just a theory of consciousness as it occurs in (say) neurotypical, adult, human beings, but instead purports to provide an account of consciousness as it might occur in infants, brain-damaged patients, non-human species and machines. This facet of IIT is a direct consequence of the fact that it is ontologically reductive, for if—as its advocates claim—consciousness *just is* integrated information, then any system with integrated information must be conscious and any conscious system must exhibit integrated information.

The Axiomatic Method

Some theories of consciousness are justified on the basis of a 'bottom-up' approach, in which one 'starts from the brain and asks how it could possibly give rise to experience' (Tononi *et al.* 2016, 450). Tononi *et al.* reject this approach on the grounds that it cannot deliver a general theory of consciousness. (In a nutshell, their worry seems to be that any attempt to construct a general theory of consciousness by looking at the neural basis of human experience must assume that the physical and behavioural correlates of human consciousness apply more widely (e.g. to other animal species and machines), and that no such assumption could be justified.) Instead, they argue, we should adopt a 'top-down' approach to consciousness. Here, one begins with the 'essential phenomenal properties of experience, or axioms, and infers postulates about the characteristics that are required of its physical substrate' (Tononi *et al.* 2016, 450). I will shortly consider the five axioms that Tononi *et al.* appeal to, but let us first reflect on the axiomatic approach itself.

There are three elements to consider in evaluating the axiomatic approach. First, there are the axioms themselves. Second, there is the relationship between the axioms and the postulates of IIT. Third, there is the question of what contribution its axiomatic foundations are meant to make to the overall epistemic status of IIT.

Let us begin with the notion of an axiom. An 'axiom', as the term is used in IIT, is a thesis about the *subjective* nature of consciousness that is *self-evidently essential* to consciousness. [Tononi and Koch (2015, 5): 'Ideally, axioms are essential (apply to all experiences), complete (include all the essential properties shared by every experience), consistent (lack contradictions) and independent (not derivable from each other).'] Axioms are not merely guiding hypotheses that might be jettisoned at a later date should they turn out to be unwarranted. Instead, they are *bedrock principles*: 'axioms are self-evident truths about consciousness—the only truths that, with Descartes, cannot be doubted and do not need proof' (Oizumi *et al.* 2014, 2). The idea seems to be that if a thesis is a genuine axiom, then its truth must be evident to any conscious creature who is able to understand it and has subjected it to serious reflection. The fact that there are reasonable individuals who, having reflected on a thesis, fail to find it compelling, not only provides evidence that it's not axiomatic but arguably makes it the case that it isn't axiomatic. After all, any number of claims about consciousness might strike someone as self-evidently true. What matters is whether the truth of the relevant thesis is regarded as self-evident within the community of consciousness researchers.

Let us turn now to the idea that the axioms are *essential* to consciousness. Here, it is important to recognize that IIT aims to provide a comprehensive theory of consciousness. Thus, in order to count as an axiom a thesis must describe a subjective feature of consciousness that applies not only to neurotypical adult members of our own species but to any possible subject of consciousness, including infants, brain-damaged individuals, non-human animals and artificially intelligent agents. As we will see, the unrestricted nature of an axiom imposes a serious burden on any attempt to establish the axiomatic status of a claim.

It is important to recognize that within IIT axioms are sharply distinguished from postulates. Postulates are defined as 'assumptions, derived from axioms, about the physical substrates of consciousness (mechanisms must have causal power, be irreducible, etc.), which can be formalized and form the basis of the mathematical framework of IIT' (Oizumi *et al.* 2014, 4). The axioms provide the starting point for IIT, but it is the postulates that provide IIT with its content. The key question is how the transition from axioms to postulates is understood within IIT.

On the most natural reading of the IIT literature, the relationship between axioms and postulates is *deductive*. On this view, the truth of the postulates can be shown to follow from the axioms with necessity in much the way in which certain geometrical theses follow from Euclid's axioms. This interpretation of IIT is suggested by multiple passages. For example, we are told that in IIT the axioms are 'formalized' into postulates (Oizumi et al. 2014, 1); that IIT 'translates' the axioms into postulates (Oizumi *et al.* 2014, 2) and that the postulates are 'derived' from the axioms (Oizumi *et al.* 2014; Tononi *et al.* 2016) states that.

Despite these passages, however, it is possible that the architects of IIT have a non-deductive conception of the relationship between the axioms and the postulates. One possibility is that the relationship between the axioms and the postulates is to be understood *abductively*—that is, it takes the form of an inference to the best explanation (Lipton 2004). On this view, the axioms are analogous to an observation (e.g. that the streets are wet), and the postulates are warranted in the way in which an abductive explanation of that observation (e.g. 'It rained last night') is warranted. Thus understood, the postulates would be taken to provide only one of several possible explanations for the axioms, and the inference from the axioms to the postulates would need to appeal to external (non-axiomatic) considerations in just the way in which the inference from 'The streets are wet' to 'It rained last night' does. (By way of contrast, no such external considerations would be needed on the deductive account.)

Although the abductive account is far more plausible than the deductive one, it is doubtful whether the abductive interpretation captures IIT as it has been presented in the literature to date. For one thing, if the inference from axioms to postulates has been understood abductively, then it is unclear why this transition has been described in terms of 'formalization', 'translation' or 'derivation'. After all, such terms are not used to describe other abductive inferences, such as the transition from 'The streets are wet' to 'It rained last night'. Second, if the relationship between axioms and postulates is understood abductively, then one would need to show not only that the postulates of IIT account for the axioms, but also that they provide a *better* account of the axioms than competing accounts do. However, the IIT literature makes no attempt to show that IIT does provide the best of the available explanation for the axioms—indeed, other possible explanations of the axioms aren't even considered.

I will leave it to the advocates of IIT to clarify the relationship between the axioms and the postulates. Here, I will assume

only that this relationship is epistemic, and that the truth of the axioms is meant to provide robust (although perhaps not conclusive) evidence for IIT (i.e. for its postulates). (A third possibility is that the axioms should be understood as playing a merely heuristic role in IIT. On this view, there would be no evidential or justificatory relationship between the axioms and the postulates. But although this is a possible interpretation of IIT, as a reading of the IIT literature it is decidedly less plausible than either the deductive or abductive interpretations.)

The final piece in this puzzle concerns the impact of the axioms on the overall epistemic status of IIT. It seems evident that the axiomatic approach is meant to establish the truth of IIT. Although the architects of IIT do appeal to other, non-axiomatic sources of evidence—such as the fact that IIT purports to explain why the cerebral cortex is associated with consciousness in a way that the cerebellum is not (Tononi *et al.* 2016, 158)—these sources of evidence are presented as having a secondary status when compared to the evidence provided by the axioms. And that attitude is entirely reasonable, given that the nature of the axiomatic approach. After all, if IIT does indeed follow (either deductively or abductively) from self-evident truths, then its epistemic credentials should be secure.

With these points in mind let us turn now to the axioms of IIT. Are they plausibly regarded as self-evident truths about the essential nature of consciousness?

The Axioms of IIT

The latest incarnation of IIT appeals to five axioms (Oizumi *et al.* 2014). This section considers each axiom in turn, asking what it means, whether it is self-evident and whether it could constrain a theory of consciousness.

The axiom of intrinsic existence

Tononi and Koch (2015) explicate the axiom of intrinsic existence as follows:

> Consciousness *exists*: my experience just is. Indeed, that my experience here and now exists—it is real or actual—is the only fact I am immediately and absolutely sure of, as Descartes realized four centuries ago. Moreover, my experience exists from its own intrinsic perspective, independent of external observers. (Tononi and Koch 2015, 5).

This passage suggests a number of claims. One claim is that consciousness exists as a genuine feature of the world—it is not an illusion, nor is it an explanatory fiction.

Is this claim axiomatic? Theorists who defend 'fictionalist' or 'illusionist' accounts of consciousness would certainly challenge it (Dennett 2016; Frankish 2016), as would those who suggest that consciousness might not be a genuine scientific kind (e.g. Allport 1988; Papineau 1993; Rey 2009; Irvine 2012). But let us grant that consciousness is real, and that this fact is self-evident. Does it follow that we have a genuine axiom here?

Perhaps, but an axiom will be useful only if it provides a substantive constraint on a theory of consciousness. What substantive constraint could this axiom impose? After all, *any* substantive theory of consciousness presupposes that consciousness is a genuine feature of the world—something that needs to be explained rather than explained away. The only accounts of consciousness that might fall foul of this constraint are certain versions of fictionalism or eliminativism, but the advocates of those views won't accept the axiom of intrinsic existence in the first place.

A second claim that is suggested by the passage reproduced above is that consciousness is an intrinsic property. On this view, an entity is conscious in virtue of the way that it itself, and nothing else, is. Although many scientists appear to regard this claim as self-evidently true, few philosophers do. According to externalist accounts of consciousness, an entity's conscious state is constitutively dependent on its history and/or relations to its environment (e.g. Dretske 1995; Hurley 1998; Lycan 2001; Byrne and Tye 2006). These 'externalist' accounts of consciousness are controversial, but they are certainly not self-evidently wrong and shouldn't be dismissed.

In sum, the so-called 'axiom of intrinsic existence' appears to be unable to provide a useful constraint on theories of consciousness. The claim that consciousness exists might indeed be axiomatic but it fails to impose a substantive constraint on theories of consciousness, while the claim that consciousness is a purely intrinsic property imposes a substantive constraint on theories of consciousness but isn't axiomatic.

The axiom of composition

Here is the axiom of composition:

> The axiom of composition states that experience is structured, being composed of several phenomenal distinctions that exist within it. For example, within an experience, I may distinguish a piano, a blue colour, a book, countless spatial locations, and so on. (Tononi *et al.* 2016, 450.)

Although this axiom is *prima facie* compelling, a number of problems emerge on closer inspection.

Recall that the axioms of consciousness capture *essential* features of consciousness. Ordinary adult human experience contains multiple contents, but it is not obvious that the same can be said of all forms of experience. What about neonatal or meditative experience? What about the experience of simple organisms or artificial agents? The advocates of IIT face some particularly challenging questions here given their endorsement of panpsychism, and the claim that 'even a binary photodiode is not completely unconscious' (Tononi 2008, 236). If extremely simple entities can be conscious, why couldn't their experiences be unstructured?

Further, it is implausible to suppose that the mere existence of phenomenal differentiation could provide a useful constraint on theories of consciousness. After all, every extant theory of consciousness recognizes that consciousness contains phenomenal differentiation of various kinds. What they disagree about is the nature of that differentiation, how it is generated and how the explanation for one kind of differentiation is related to the explanation of other forms of phenomenal differentiation. But the axiom of composition says only that consciousness *is* differentiated, and this claim has no bearing on those debates.

Of course, one *could* use claims about phenomenal differentiation to impose a substantive constraint on a theory of consciousness. This approach has been employed by the advocates of the intermediate-level theory of consciousness, who argue that the contents of consciousness are restricted to 'intermediate-level' representations (e.g. Jackendoff 1987; Prinz 2011). Whether or not that view is compelling (see Bayne 2009; Hawley and Macpherson 2011; Kemmerer 2015; McClelland and Bayne 2016 for contrary views), the key point is that this constraint is very different from the constraint to which the architects of IIT appeal, for it is not advanced as a self-evident truth about all

possible experience but as an empirical claim about the structure of human experience.

In sum, ordinary human experience certainly involves phenomenal differentiation, but it is doubtful whether 'phenomenal composition' is an essential feature of consciousness. Even if it were, that claim would not place a substantive constraint on a general theory of consciousness.

The axiom of information

The axiom of information is one of the most puzzling of IIT's five axioms. Here is how Tononi and Koch introduce it:

> Consciousness is *specific*: each experience is *the particular way* it is—it is composed of a specific set of specific phenomenal distinctions—thereby differing from other possible experiences (*differentiation*). Thus, an experience of pure darkness and silence is what it is because, among other things, it is not filled with light and sound, colours and shapes, there are no books, no blue books and so on. And being that way, it necessarily differs from a large number of alternative experiences I could have. Just consider all the frames of all possible movies: the associated visual percepts are but a small subset of all possible experiences. (Tononi and Koch 2015, 6; emphasis in original)

It is no doubt true that each experience 'is the particular way that it is', but the same can be said of any phenomenon (every toaster is the particular way that it is; every hamster is the particular way that it is; and so on). Tautologies may be self-evident but they illuminate nothing.

Perhaps progress can be made by considering Tononi's (2008) discussion of a photodiode. A photodiode registers the difference between a screen being off and it being switched on, but—Tononi claims—it lacks the capacity to experience light and dark. [But note that Tononi also claims that 'even a binary photodiode is not completely unconscious' (Tononi 2008, 236; see also Oizumi *et al.* 2014, 19).] What, he asks, is the key difference between organisms like us and the photodiode? The answer, he says, concerns the range of discriminations that you can make as opposed to the range of discriminations that the diode can make:

> When the blank screen [of the photodiode] turns on, the mechanism in the photodiode tells the detector that the current from the sensor is above rather than below the threshold, so it reports 'light.' In performing this discrimination between two alternatives, the detector in the photodiode generates log2 (2) _ 1 bit of information. When you see the blank screen turn on, on the other hand the situation is quite different. Though you may think you are performing the same discrimination between light and dark as the photodiode, you are in fact discriminating among a much larger number of alternatives, thereby generating many more bits of information. This is easy to see. Just imagine that, instead of turning light and dark, the screen were to turn red, then green, then blue, and then display, one after the other, every frame from every movie that was ever produced. The photodiode, inevitably, would go on signaling whether the amount of light for each frame is above or below its threshold: to a photodiode, things can only be one of two ways, so when it reports 'light', it really means just 'this way' versus 'that way'. For you, however, 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). . . . According to the IIT, it is all this added meaning, provided implicitly by how we discriminate pure light from all these alternatives, that increases the level of consciousness. . . . [IIT] says that the more specifically one's mechanisms discriminate between what pure light is and what it is not (the more they specify what light means), the more one is conscious of it. (Tononi 2008, 217–8)

A number of claims are suggested by this passage. One claim is this:

> INFORMATION^1: A creature's level of consciousness is a function of the range of discriminations that it can make.

Whatever plausibility INFORMATION^1 might have in its own right, it doesn't provide us with a possible interpretation of the axiom of information, for INFORMATION^1 is a claim about what it is for an entity to have a certain *level* of consciousness whereas what we need here is a claim about the essential features of consciousness *per se*. So, we will leave INFORMATION^1 to one side.

A second interpretation of the axiom of information is this:

> INFORMATION^2: The capacity to have a conscious content (e.g. that the light is on) requires the capacity to have a range of conscious contents.

INFORMATION^2 captures the idea that the contents of consciousness are holistic. They cannot exist as singletons but occur only in the context of bundles of such capacities.

The idea that conscious capacities occur only as bundles is certainly attractive. However, there are serious difficulties involved in spelling this idea out in a manner that is both precise and plausible. The main problem with INFORMATION^2 is that in order to generate a truth-evaluable thesis more must be said about the 'range' of capacities that is required for consciousness of any one content. Tononi himself seems to think that a very broad range of capacities is required: 'For you, the 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).' Taken literally, these comments entail an implausible form of holism about consciousness, for one can experience a screen as light without having the capacity to experience 'a red screen, a green screen, a blue screen, this movie frame, that movie frame, and so on for every movie frame'. After all, individuals suffering from achromatopsia have lost the capacity to experience colours, but they retain the capacity to experience luminance (motion, figure, etc.). If INFORMATION^2 is to be at all plausible then the range of representational capacities that it appeals to must be constrained, but it is unclear what form such constraints might take or how they might be motivated. One might also ask whether INFORMATION^2 is consistent with the commitment to panpsychism that the advocates of IIT share. There certainly seems to be a tension between thinking that (say) a diode can have some level of consciousness and thinking that being conscious requires the capacity to have a wide range of contents in consciousness.

A third interpretation of this axiom views it as a form of conceptual role semantics, an account that equates the content of a representation with its causal, functional or inferential role (e.g. Loar 1981; Harman 1982; Block 1986). Here is one version of this idea:

> INFORMATION^3: The content of any one state of a conscious system is determined solely by the causal, inferential or functional relations that it bears to every other state of the system.

Whether or not INFORMATION^3 captures what the advocates of IIT mean by 'the axiom of information', there is good reason to think that they would endorse it (or at least something very much like it). Is INFORMATION^3 axiomatic?

That is highly unlikely, for even if INFORMATION[3] is true its truth is hardly self-evident. Conceptual role semantics is only one of many accounts of content, and it is very far from being the most influential account of content (see Loewer 1997). Its minority status is well justified, for there are many problems with it. One problem is that it entails that no two people can share any one content (or concept) unless they share *all* of their contents (or concepts). Not only is that implication deeply implausible in its own right, it is also at odds with the practice of consciousness science, which assumes that individuals with very different experiences can have certain contents in common.

In sum, the interpretations of the axiom of information that are plausibly regarded as axiomatic fail to impose a substantive constraint on consciousness, whereas the interpretations of this axiom that impose a substantive constraint on consciousness are not plausibly regarded as axiomatic.

The axiom of integration

Here is the axiom of integration:

> The axiom of integration states that experience is unitary, meaning that it is composed of a set of phenomenal distinctions, bound together in various ways, that is irreducible to non-interdependent subsets. (Tononi *et al.* 2016, 452)

The architects of IIT do not provide an analysis of what it is for experience to be unitary, nor do they say what it is for consciousness to be 'irreducible to non-interdependent subsets'. They do, however, provide examples of the kind of unity in which they are interested: the experience of a whole visual scene cannot be subdivided into independent experiences of the left and right sides of the visual field (Tononi *et al.* 2016, 452); the experience of the word 'SONO' written in the middle of a blank page is irreducible to an experience of the word 'SO' on one's right and an experience of the word 'NO' on one's left (Oizumi *et al.* 2014, 3) and the experience of seeing a red triangle is irreducible to seeing a triangle and an experience of seeing redness (Oizumi *et al.* 2014, 3).

These examples suggest that the axiom of integration concerns *representational unity* (Bayne and Chalmers 2003; Bayne 2010). Two experiences (e^1, e^2) are representationally unified if (and only if) they occur as components of a complex experience whose content cannot be identified with the conjunction of the contents of e^1 and e^2. For example, the experience of a word as a token of 'SONO' cannot be identified with a conjunction of an experience of 'SO' and an experience of 'NO', for it is coherent to suppose that one could have experiences of 'SO' and 'NO' without representing the whole stimulus as the word 'SONO'.

In order for representational unity to form the basis of an axiom, it must figure in a truth-evaluable thesis. The following is an example of what such a thesis might look like:

> INTEGRATION[1]: Necessarily, any two experiences that are had by the same subject of experience at the same time are representationally unified with each other.

Is INTEGRATION[1] axiomatic? That seems unlikely. Although representational unity is indeed a common feature of consciousness, there are pathologies of consciousness in which it breaks down. For example, in associative agnosia, patients experience the individual features and parts of objects but fail to synthesize those features and parts into an integrated percept (Humphreys and Riddoch 1987; Farah 2004). There are also counter-examples to INTEGRATION[1] in ordinary experience. For example, one can hear birdsong and feel a dull pain in one's calves without these two experiences being representationally unified with each other. Nor is there anything unusual about these experiences: many of the experiences that co-occur within one's stream of consciousness are not representationally unified with each other. And even if the contents of *human* experience *were* always mutually representationally unified, what reason is there to think that this kind of unity is an *essential* feature of consciousness?

A second interpretation of the axiom of integration appeals to what Dainton (2000) calls 'gestalt unity'. Two experiences are gestalt unified when they require each other's presence even though there are no entailment relations between their contents. Thus, an experience of birdsong would be gestalt unified with a pain experience if and only if it is the case that the existence of the pain demands the existence of the birdsong and vice-versa, even though the content of neither experience entails the content of the other.

Gestalt unity suggests the following version of the axiom of integration:

> INTEGRATION[2]: Every phenomenal content ('distinction') that occurs within a subject's overall conscious state is gestalt unified with every other phenomenal content ('distinction') that occurs within it.

Although INTEGRATION[2] may do a better job of capturing this axiom than INTEGRATION[1] does, it too is implausible. The problem with INTEGRATION[2] is that it makes the relations between any two unified experiences necessary, and that entails that no two individuals could share *any* kind of experience unless they share *all* of their experiences. This claim is not only *prima facie* implausible, it is also at odds with the practice of consciousness science. Researchers studying (e.g.) shape perception invariably assume that two subjects can experience the same shape even if they are in very different mood states (for example). Gestalt unity is at best a rare phenomenon: it is not even essential to human experience let alone all possible forms of experience.

A third interpretation of the axiom of integration focuses on the relationship between particular experiences. Distinguish two views of the structure of consciousness: phenomenal atomism and phenomenal holism (Bayne 2010; Bayne 2014). According to the phenomenal atomist, complex experiences (such as the one that captures what it is like to be you right now) are constructed from sets of simpler experiences. To take a very crude example, your current experience might be built up out of (say) a visual experience of this article, an experience of one's body as having a certain position in space and a background mood phenomenology. Atomists treat these experiences (or at any rate something akin to them) as the building blocks of complex experiences. They take a subject's overall phenomenology to be generated by 'gluing' these building blocks together in various ways, perhaps by means of relations of 'co-consciousness' [also known as 'phenomenal unity' (Bayne 2001)].

Holists, in contrast, regard the subject's overall conscious experience as the basic unit of consciousness. On this view, although there is phenomenal differentiation within one's overall conscious experience (e.g. a visual experience, an experience of one's body, a mood experience), this differentiation doesn't demarcate independent units of consciousness, and thus there is no need for any kind of 'glue' that might bind these units together.

Whether or not phenomenal holism captures what the architects of IIT have in mind, it certainly represents a possible interpretation of the claim that consciousness is 'irreducible to

non-interdependent subsets of experience'. In light of this, we might consider a third interpretation of this axiom:

> INTEGRATION^3: No experience can be built up out of simpler experiences.

Although I myself have argued in favour of a restricted form of phenomenal holism (Bayne 2010; Bayne 2014), it seems unlikely to me that INTEGRATION^3 could qualify as an axiom of consciousness.

There are two problems with treating INTEGRATION^3 as axiomatic, each of them fatal. Firstly, the contrast between phenomenal holism and phenomenal atomism is not manifest from the subjective point of view, for one has no first-person access to the underlying causal structure of consciousness. The atomism and the holist can each allow that ordinary human experience includes a 'number of phenomenal distinctions that are bound together in various ways', it's just that they give different explanations for this fact. The atomist regards at least some of these distinctions as involving distinct experiences that are bound together by a genuine relation, whereas the holist regards these distinctions as merely differences in content and the binding relation as merely nominal. Second, even if phenomenal holism is true of us, there is no reason to think that it captures an essential feature of consciousness.

Again, we have failed to find a thesis about the essential nature of consciousness that is both self-evidently true and that also provides a substantive constraint on theories of consciousness.

The axiom of exclusion

The fifth and final axiom is the axiom of exclusion:

> Consciousness is definite, in content and spatio-temporal grain: each experience has the set of phenomenal distinctions it has, neither less (a subset) nor more (a superset), and it flows at the speed it flows, neither faster nor slower. Thus, the experience I am having is of seeing a body on a bed in a bedroom, a bookcase with books, one of which is a blue book, but I am not having an experience with less content—say, one lacking the phenomenal distinction blue/not blue, or coloured/not coloured; nor am I having an experience with more content—say, one endowed with the additional phenomenal distinction high/low blood pressure. Similarly, my experience flows at a particular speed—each experience encompassing a hundred milliseconds or so—but I am not having experience that encompasses just a few milliseconds or instead minutes or hours. (Tononi and Koch 2015, 6)

Let us begin with the claim that the *contents* of consciousness are 'definite'. The central question here is whether Tononi *et al.* take the 'definiteness' of conscious content to rule out the possibility of vagueness.

Presentations of the axiom of exclusion contain no explicit commitment to the idea that vague contents are impossible. Instead, their only explicit commitment is to the idea that certain contents *are* determinately present in one's experience (e.g. seeing a body on a bed in a bedroom) and certain contents are determinately *not* present in one's experience (e.g. having high/low blood pressure). That claim, of course, is consistent with the possibility that certain contents that are neither determinately present nor determinately not present. (Compare: some people are determinately short, some people are determinately tall and some people are neither determinately short nor determinately tall.) But if the axiom of exclusion is understood in this manner, then it is unclear what constraints it places on a theory of consciousness, for every theory of consciousness holds that some contents are determinately present in experience and some contents are determinately not present in experience.

But let us suppose, if only for the sake of argument, that the axiom of exclusion *is* intended to rule out the possibility of vagueness in the contents of consciousness. Would that claim be axiomatic? No. Although some theorists have argued on *a priori* grounds that consciousness is not vague (e.g. Antony 2006, 2008; Simon 2017), that view is controversial and is rejected by a number of other theorists (e.g. Papineau 1993; Tye 1996). Moreover, even if consciousness itself cannot be vague, it doesn't follow that the *contents* of consciousness cannot be vague. And indeed, there is every reason to think that consciousness can have vague contents. Consider an experience of a striped tiger. Does this experience represent the tiger as having a precise number of visible stripes—say, 78? Possibly; but it is *prima facie* more plausible to suppose that one's visual experience contains some indeterminacy in how many stripes it represents the tiger as having. It is doubtful whether there is any plausible version of the axiom of exclusion that focuses on the contents of consciousness.

Let us turn now to the claim that the 'duration of the instant of consciousness is also definite, ranging from a few tens of milliseconds to a few hundred milliseconds'. The temporal structure of consciousness certainly provides the science of consciousness with an important source of constraints for theory building. But is it plausible to view those constraints through the lens of the axiomatic approach? I don't think so.

There are two main problems here. The first concerns the content of the alleged constraint. Suppose that an 'instant of consciousness' (whatever exactly that means) has a particular duration ('*D*') irrespective of the kind of creature in question or its state of consciousness. If that were the case, then we might begin to suspect that *D* was an essential feature of consciousness, and something that ought to be regarded as such by any plausible theory of consciousness. But we have no reason to think that there *is* a particular duration that characterizes consciousness in all kinds of creatures and all kinds of conscious states. Indeed, immediately after claiming that the duration of an instant of consciousness is 'definite', Tononi *et al.* go on to say that this duration ranges from 'a few tens of milliseconds to a few hundred milliseconds'. In other words, the architects of IIT themselves regard the duration of consciousness as variable. But if that is the case, then what kind of constraint could the axiom of exclusion place on a theory of consciousness?

The second problem here concerns what is meant by 'an instant of consciousness', and whether this is something to which we might have reliable, first-person, access. One problem here is that there are good reasons to think that we have direct access only to the temporal relations between the *contents* of consciousness, and that first-person access to the temporal properties of conscious experiences themselves is indirect and inferential, mediated by our access to the temporal relations between their contents (see e.g. Dennett 1991; Dennett and Kinsbourne 1992). Further, our capacity to determine the fine-grained temporal relations between events is notoriously poor, and thus there is no reason to regard our judgments about the duration of an instant of consciousness as 'self-evident' or 'indubitable' (see e.g. Stone *et al.* 2001; Haggard *et al.* 2002; Spence and Parise 2010).

Let us take stock. It has proven very difficult to identify theses that could play the role that IIT requires of its axioms. Some theses that are advanced as axioms arguably qualify as self-evident truths about the essential features of consciousness but they fail to provide substantive constraints on a theory of consciousness,

whereas other theses might provide substantive constraints on a theory of consciousness but are not plausibly regarded as self-evident truths about the essential features of consciousness. In short, the axiomatic foundations of IIT are shaky.

From the Axiomatic Approach to the Natural Kind Approach

Let us take a step back from the details of IIT itself and consider the viability of the axiomatic approach itself. There are good reasons to think that the axiomatic method is not well-suited to the study of consciousness. Axiomatic methods are most closely associated with mathematics and logic, and one will not find any mention of them in accounts of explanation in the mechanical or life sciences (Cummins 1985; Bechtel 2007; Craver 2007). Thus, to the extent that one is attracted to the idea that the study of consciousness has its natural home in neuroscience or psychology, one ought to be sceptical of the axiomatic approach.

Second, it is debatable whether there *are* any (non-trivial) essential, subjective properties of consciousness (over and above the fact that there is 'something it is like' to be in a specific type of conscious state). And even if consciousness does have essential features, it is not clear that we have the capacity to identify them. After all, the only form of consciousness to which we have direct access is our own. We don't have direct access to all forms of human consciousness, let alone the kind of consciousness that characterizes non-human species or machines.

But if the axiomatic approach is ill-suited to the science of consciousness how then *should* we proceed? This is a tough question, and a through response to it deserves a paper of its own. Here, I have space only to sketch one alternative to the axiomatic approach—the *natural kind* approach (Shea and Bayne 2010; Shea 2012; see also Seth *et al.* 2008; Peterson 2016).

The natural kind approach proceeds by treating consciousness as a natural kind, akin to gold, water or hepatitis. One begins with the various signs (markers; symptoms) with which it is associated and then attempts to discover how those signs cluster together. Having identified such clusters, one then searches for the underlying mechanism(s) that accounts for them. Once one has identified these mechanisms, one is then able to determine the distribution of consciousness and perhaps provide some kind of explanation of it. In the case of hepatitis, pursuing the natural kind approach led to the identification of certain viruses, the presence of which explains why the signs that were pre-theoretically associated with hepatitis cluster together in the ways that they do (Seeff 2009). Having found these viruses, we can now use tests for their presence to determine the distribution of hepatitis in a population.

There are a number of important points of contrast between the axiomatic approach and the natural kind approach. Firstly, the axiomatic approach is concerned only with features of the target phenomenon that are (putatively) essential to it, whereas the natural kind approach is concerned both with (putatively) essential features and with features that are merely associated with the target phenomenon. Consider those domains in which the natural kind approach has been very successful, such as the study of disease. Many of the signs and symptoms of a disease are not essential to it, and yet the interrogation of such signs and symptoms often makes an invaluable contribution towards understanding its nature. Indeed, the fact that there are certain conditions in which the typical signs of consciousness appear to *dissociate* from each other can itself be a useful data point when it comes to the search for underlying mechanisms. Secondly, the natural kind approach does not restrict itself to the phenomenological dimensions of consciousness in the way that the axiomatic approach does, but considers also the relationship between consciousness and other psychological states and capacities, such as attention, working memory, introspective accessibility and the intentional control of behaviour.

Rather than begin with the search for self-evident truths about the essential, phenomenological features of consciousness (as the axiomatic approach does), the natural kind approach recommends that we begin by looking for clusters between the various signs of consciousness and then try to explain why those clusters obtain. Whether or not the natural kind approach is able to address the hard problem of consciousness (Chalmers 1996), it is a recognized approach to explanation in the biological sciences, and would appear to fit the science of consciousness far better than the axiomatic approach does.

Would it be possible to develop IIT within the framework of the natural kind approach? I don't see why not. Even if the so-called 'axioms' that Tononi *et al.* appeal to fail to qualify as genuine axioms, the features of consciousness to which they appeal could still play a vital role within the context of the natural kind approach. For example, any account of consciousness needs to explain why human consciousness is typically unified in the various ways that it is. The fact that IIT appears to have a very natural explanation for this fact is surely a mark in its favour, especially when one considers that many theories of consciousness make no attempt at all to account for it.

Funding

This article was written with the help of an *Australian Research Council Future Fellowship* (FT 150100266) and the support of the *Canadian Institute for Advanced Research* (CIFAR). I have also benefitted from comments by David Chalmers, Daniel Mathews, Anil Seth, Giulio Tononi, Nao Tsuchiya, and three referees for this journal. Supplementary data is available on request.

References

Allport A. What concept of consciousness? In Marcel A, Bisiach E (eds), *Consciousness in Contemporary Science*. Oxford: OUP, 1988, 159–82.

Antony M. Vagueness and the metaphysics of consciousness. *Philos Stud* 2006;**128**:515–38.

Antony M. Are our concepts conscious state and conscious creature vague? *Erkenntnis* 2008;**68**:239–63.

Bayne T. Co-consciousness. *J Consciousness Studies* 2001;**8**:79–92.

Bayne T. The multisensory nature of perceptual consciousness. In Bennett W and Hill C (eds), *Sensory Integration and the Unity of Consciousness*. Cambridge, MA: MIT Press, 2014, 15–36.

Bayne T. Perception and the reach of phenomenal content. *Philos Q* 2009;**59**:385–404.

Bayne T. *The Unity of Consciousness*. Oxford: Oxford University Press, 2010.

Bayne T, Chalmers D. What is the unity of consciousness? In Cleeremans A (ed.), *The Unity of Consciousness*. Oxford: Oxford University Press, 2003, 23–58.

Bechtel W. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New Jersey: Lawrence Erlbaum, 2007.

Block N. Advertisement for a semantics in psychology. In French P, Uehling T, Wettstein H (eds), *Midwest Studies in Philosophy, Vol. 10. Studies in the Philosophy of Mind*. Minneapolis, MN: University of Minnesota Press, 1986, 615–78.

Byrne A, Tye M. Qualia ain't in the head. *Nous* 2006;**40**:241–55.
Casali AG, Gosseries O, Rosanova M, *et al*. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 2013;**5**:198ra105.
Casarotto S, Comanducci A, Mario R, *et al*. Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann Neurol* 2016;**80**:718–29.
Chalmers D. *The Conscious Mind*. Oxford: Oxford University Press, 1996.
Craver C. *Explaining the Brain*. Oxford: OUP, 2007.
Cummins R. *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press, 1985.
Dainton B. *Stream of Consciousness: Unity and Continuity in Conscious Experience*. London: Routledge, 2000.
Dennett D. *Consciousness Explained*. Boston, MA: Brown and Little, 1991.
Dennett D. Illusionism as the obvious default theory of consciousness. *J Conscious Stud* 2016;**23**:65–72.
Dennett D, Kinsbourne M. Time and the observer: there where and when of consciousness in the brain. *Behav Brain Sci* 1992; **15**:183–201.
Dretske F. *Naturalizing the Mind*. Cambridge, MA: MIT Press, 1995.
Farah MJ. *Visual Agnosia, 2nd edn*. Cambridge, MA: MIT Press, 2004.
Frankish K. Illusionism as a theory of consciousness. *J Conscious Stud* 2016;**23**:11–39.
Haggard P, Clark S, Kalogeras J. Voluntary action and conscious awareness. *Nat Neurosci* 2002;**5**:382–5.
Harman G. Conceptual role semantics. *Notre Dame J Formal Logic* 1982;**23**:242–56.
Hawley K, Macpherson F. *The Admissible Contents of Experience*. Oxford: Wiley-Blackwell, 2011.
Humphreys GW, Riddoch MJ. *To See But Not to See*. Hillsdale: Erlbaum, 1987.
Hurley S. *Consciousness in Action*. Cambridge, MA: Harvard University Press, 1998.
Irvine E. *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*. Dordrecht: Springer, 2012.
Jackendoff R. *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press, 1987.
Kemmerer D. Are we ever aware of concepts? A critical question for the global neuronal workspace, integrated information, and attended intermediate-level representation theories of consciousness. *Neurosci Conscious* 2015; doi: 10.1093/nc/niv006.
Lipton P. *Inference to the Best Explanation, 2nd edn*. London: Routledge, 2004.
Loar B. *Mind and Meaning*. Cambridge: Cambridge University Press, 1981.
Loewer B. A guide to naturalizing semantics. In Wright C, Hale B (eds), *A Companion to the Philosophy of Language*. Blackwell: Oxford, 1997, 108–26.
Lycan W. The case for phenomenal externalism. *Philos Perspect* 2001;**15**:17–35.
McClelland T, Bayne T. Concepts, contents, and consciousness. *Neurosci Conscious* 2016;**1**:1–9.
Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comp Biol* 2014;**10**:e1003588.
Papineau D. *Philosophical Naturalism*. Oxford: Blackwell, 1993.
Peterson A. Consilience, clinical validation, and global disorders of consciousness. *Neurosci Conscious* 2016; doi: 10.1093/nc/niw011
Prinz J. The sensory basis of cognitive phenomenology. In Bayne T, Montague M (eds), *Cognitive Phenomenology*. Oxford: OUP, 2011, 174–96.
Rey G. 'Elimativism.' In Bayne T, Wilken P, Cleeremans A (eds), *The Oxford Companion to Consciousness*. Oxford: OUP, 2009, 252–3.
Seeff LB. The history of the "natural history" of hepatitis C (1968–2009). *Liver Int* 2009;**29**:89–99.
Seth A, Dienes Z, Cleeremans A, *et al*. Measuring consciousness: relating behavioural and neurophysiological measures. *Trends Cogn Sci* 2008;**12**:314–21.
Shea N. Methodological encounters with the phenomenal kind. *Philos Phenomenolog Res* 2012;**84**:307–44.
Shea N, Bayne T. The vegetative state and the science of consciousness. *Br J Philos Sci* 2010;**61**:459–84.
Simon J. Vagueness and zombies: why 'phenomenally conscious' has no borderline cases. *Philos Stud* 2017;**174**: 2105–23.
Spence C, Parise C. Prior entry. *Conscious Cogn* 2010;**19**:364–79.
Stone JV, Hunkin NM, Porrill J, *et al*. When is now? Perception of simultaneity. *Proc R Soc (B)* 2001;**268**:31–8.
Tononi G. Consciousness and integrated information: a provisional Manifesto. *Biol Bull* 2008;**215**:216–42.
Tononi G, Boly M, Massamini M, *et al*. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.
Tononi G, Koch C. Consciousness: here, there and everywhere? *Philos Trans R Soc B* 2015;**370**:20140167.
Tye M. Is consciousness vague or arbitrary? *Philos Phenomenolog Res* 1996;**56**:679–85.