2019

Interpretation-Neutral Integrated Information Theory

Kelvin J. McQueen

Interpretation-Neutral Integrated Information Theory*

Kelvin J. McQueen†

December 23, 2018

Abstract

The integrated information theory (IIT) is a theory of consciousness that was originally formulated, and is standardly still expressed, in terms of controversial interpretations of its own ontological and epistemological basis. These form the *orthodox* interpretation of IIT. The orthodox epistemological interpretation is the axiomatic method, whereby IIT is ultimately derived from, justified by, and beholden to, a set of phenomenological axioms. The orthodox ontological interpretation is panpsychism, according to which consciousness is fundamental, intrinsic, and pervasive. In this paper it is argued that both components of the orthodox interpretation should be rejected. But IIT should not be rejected since an interpretation-neutral formulation is available. After explaining the neutral formulation, more plausible non-axiomatic epistemologies are defended. The neutral formulation is then shown to be consistent with various contemporary physicalist ontologies of consciousness, including the phenomenal concepts strategy, representationalism, and even illusionism. Along the way, instructive connections between interpretations of IIT and interpretations of quantum mechanics, are noted.

Contents

*This is a preprint. The final and definitive version will be published in the *Journal of Consciousness Studies*.
†Department of Philosophy, Chapman University, CA 92866, United States

1 Theory and Interpretation

Scientific theories are often framed in terms of the philosophical assumptions of their founders. Perhaps the most striking example is quantum theory, which was framed, throughout much of the twentieth century, in terms of the philosophical assumptions of Danish physicist Niels Bohr. This expression of quantum theory came to be known as the Copenhagen interpretation and is commonly referred to as the *orthodox interpretation of quantum theory.*[1]

A more recent example from neuroscience is the integrated information theory of consciousness (IIT). IIT is almost invariably framed in terms of the substantive philosophical assumptions of its founder, Guilio Tononi.[2] I will therefore refer to these assumptions as the *orthodox interpretation of IIT*. When IIT is expressed in terms of these assumptions I will refer to it as *Orthodox-IIT*.

The philosophical assumptions that make up an interpretation can be distinguished into two kinds, ontological and epistemological. *Ontological interpretations* are claims about what exists (if anything) according to the theory, or what reality must be like (if anything) given the predictive success of the theory. *Epistemological interpretations* state what kind of evidence supports the theory and how.

Contemporary interpretations of quantum theory are ontological but not epistemological.[3] For it is relatively clear what the predictions of quantum theory are and we have been able to confirm those predictions with modern technology. But in the case of IIT, it is relatively unclear what its predictions are, and insofar as we can clarify them, current technology struggles to test them. This not only raises the question of why we should believe IIT. It raises the question of whether IIT can even be considered scientific. Epistemological interpretations of IIT attempt to answer these questions.

The foundations of physics is the branch of physics that evaluates interpretations of quantum theory. Since at least the 1950's it has been busy teasing apart the interpretation-neutral quantum theory from its orthodox interpretation, critically examining the orthodox interpretation, and formulating and evaluating alternative interpretations.[4] As a result, our understanding of quantum theory has significantly progressed. This progress is inspiring the present project, which aims to do the same for IIT. The basic idea behind this paper is that if quantum theory can advance by extracting its core structure from its problematic orthodox interpretation, then so too can IIT.

Interpretation-neutral integrated information theory (Neutral-IIT) is intended to be a "bare-bones" formulation that expresses only what is essential for an experimental neuroscientist to apply the theory in practice, while minimizing (or ideally, completely removing) controversial philosophical assumptions.[5] There are many reasons for defining Neutral-IIT, isolating and criticizing its orthodox interpretation, and formulating new, more plausible interpretations. One is to combat much of the undeserved negative attention that IIT has received. There are now many published criticisms of IIT. But while they do seriously challenge Orthodox-IIT, they typically fail to address Neutral-IIT. The neutral formulation opens IIT up to a broader audience who might otherwise reject it due to their philosophy of mind.

[1] See e.g. Griffith (2018, p5) and Albert (1992, p17).

[2] See Tononi (2004, 2008, 2012, 2017a, 2017b), Oizumi et. al (2014), Tononi & Koch (2015), and Tononi et. al. (2016).

[3] In quantum theory, there is a distinction between *psi-epistemic* and *psi-ontic* interpretations. These interpretations agree on what evidence supports quantum theory but give alternative accounts of the reality described by quantum theory. Hence, they are both ontological but not epistemological interpretations, see e.g. Aaronson et. al. (2013).

[4] This author's recent attempts to contribute to this field can be found in McQueen (2015) and McQueen and Vaidman (2018).

[5] For a quantum analogue of this project, see Wallace (forthcoming).

Here is the structure of the paper. Section 2 describes the mathematical and empirical content of Neutral-IIT. Section 3 describes epistemological interpretations. I begin by describing the orthodox epistemological interpretation, the axiomatic method. I then explain why the axiomatic method cannot work, and should be abandoned (Sec. 3.1). I then consider two more plausible epistemological interpretations. The first denies that the axioms have an axiomatic status, instead treating them as malleable postulates that might help to abductively confirm IIT (Sec 3.2). The second denies that the axioms play any role in justifying the theory (Sec. 3.3).

Section 4 describes ontological interpretations. I describe the orthodox ontological interpretation before raising several objections to it (Sec. 4.1). I then show that contemporary philosophy of mind provides a number of resources for developing more plausible ontologies that are immune to the objections that befall Orthodox-IIT. I describe three physicalist interpretations, based on the phenomenal concepts strategy (Sec. 4.2), representationalism (Sec. 4.3), and finally, illusionism (Sec. 4.4). I leave it to the reader to decide which is the superior physicalist interpretation of IIT.

2 Interpretation-Neutral Integrated Information Theory

Integrated information theory (IIT) is presented as a theory of *phenomenal* consciousness. A subject is phenomenally conscious if and only if there is something it is like to be that subject. A mental state is phenomenally conscious if and only if there is something it is like to be in that state. In addition to this traditional definition (Nagel (1974)), Tononi offers an instructive operational definition: phenomenal consciousness is what one loses when one falls into dreamless sleep and then regains when one either wakes up or starts dreaming (Tononi et. al. (2016, p450)). Finally, there is a dialectical definition: phenomenal consciousness is what's being debated in debates over the hard problem of consciousness (Chalmers 1995a). In what follows I refer to phenomenal consciousness just as *consciousness* or *experience*.

Consciousness seems like a private phenomenon: while I can be certain about my own consciousness, I cannot be certain about the consciousness of others. Nonetheless, consciousness manifests itself in the form of observable symptoms of consciousness. The most obvious of these are phenomenal reports, whereby one describes one's consciousness. In practice, such reports are used to find the neural correlate of consciousness.

IIT claims to have identified the neural correlate of consciousness. According to IIT, it is integrated information. Information is not like information in a book, it is an objective measurable property of physical systems. When it takes a certain form, it can be said to be integrated. I define these notions more formally below.

IIT has three core applications. It uses facts about the integrated information in a system to determine (i) whether that system is conscious, (ii) to what extent it is conscious, and (iii) the qualitative character (or qualia) of that system's experience. That's the basic idea. Let us now consider how we can present this theory in detail, but in a way that strips it of its controversial philosophical assumptions.

If the subject matter of IIT is consciousness, then the project of formulating an absolutely philosophically neutral version of IIT might seem like a non-starter. For there are many prominent philosophers who claim that consciousness does not even exist, that it is an illusion! These are the illusionists.[6]

Given illusionism, there are two options. We could concede that an absolutely neutral formulation of IIT is impossible, formulate Neutral-IIT as a theory of phenomenal consciousness that is inconsistent with illusionism, and from there try to make it as philosophically neutral as

[6]For example, see Dennett (1991, 1998, 2016) and Frankish (2016).

possible. Alternatively, we could try to find a way to define Neutral-IIT so that it is consistent with both illusionism and realism about consciousness.

Here I take the second option. The main reason is that the core structure of IIT should be open to illusionists. To see this, imagine that in the future we find out that our judgments about phenomenal consciousness seem to always strongly correlate with information integration in corresponding brain states. It is not clear that this would refute illusionism. In fact, illusionists might want to say that information integration plays some important role in creating the illusion (McQueen, forthcoming).[7]

To formulate Neutral-IIT so that it is neutral between illusionism and realism about consciousness, we must find some phenomena in the vicinity of consciousness that both parties agree upon. The empirical content (the predictions) of Neutral-IIT may then concern correlations between information integration and the agreed upon phenomena. The natural choice would be uncontroversial observable symptoms of consciousness such as phenomenal reports. Phenomenal reports are reports of how things consciously seem to subjects and how their experience feels to them, etc. What Dennett (2003) calls the subject's heterophenomenological world.

Neutral-IIT is therefore just a framework for helping researchers find the neural correlates of consciousness (or its observable symptoms). It provides an account of what to look for: integrated information in brain regions that correlate with the observable symptoms. And it provides the means for making more precise predictions and testing them (discussed below). The framework does not answer the question of why these correlations exist (the hard problem), and is consistent with answers that postulate the reality of consciousness as well as those that do not.

To define Neutral-IIT, we need two ingredients. First, we need a mathematical formalism that tells us how to calculate information integration. This is the *mathematical content* of the theory. Second, we need a clear set of predictions that correlate information integration to observable symptoms of consciousness. This is the *empirical content* of the theory. I will take each of these in turn.

2.1 The mathematical content

There are three crucial notions that require formal definitions. The first two are *information* and *integration* (or ϕ). The third has gone under various labels, and will here be referred to as *Q-shape*. A Q-shape is an abstract structural property of integrated information. Under Orthodox-IIT, for every qualitative character of experience (quale), there is a distinctive Q-shape that determines that quale (Balduzzi & Tononi (2009)).

The mathematical formalism of the bare theory is just the mathematical formalism of IIT. And so here, I will only explain the technical details that will be relevant to the subsequent discussion. We will need a technical definition of information. Integration and Q-shape will play less of a role in subsequent discussion and will receive only simple intuitive definitions.(Readers already familiar with the mathematical formalism may proceed to section 2.2.)

In IIT, *information* is a measure of the extent to which the present state of a system constrains that system's potential past and future states. We may therefore speak of the *cause information* in the system, which is a measure of the extent to which the present state of the system constrains its immediate past state, and the *effect information* in the system, which is a measure of the extent to which the present state of the system constrains its immediate future

[7]There is an important parallel here with quantum theory. In addition to realist ("psi-ontic") interpretations of quantum theory, there are also anti-realist ("psi-epistemic") interpretations which deny the existence of a quantum physical reality, e.g. Caves et. al. (2002). If we allow quantum interpretations that deny that the quantum formalism describes a real microphysical reality, then we should also allow interpretations of IIT that deny that the IIT formalism describes the reality of consciousness.

state. The information (per se) is defined as the minimum of the cause information and the effect information.[8]

Let's consider how we measure the *cause* information in a system (the same principles apply to the measurement of the effect information). First, we need to identify the state space of the system, i.e. all of its possible configurations. We must then assign two probability distributions to this state space, an *a priori* distribution, and an *a posteriori* distribution.

The *a priori* probability distribution is relatively trivial as it assigns equal probability to each possible state in the state space. It can be thought of as the probability that the system was, in the previous moment, in one of its possible states, given that we know *nothing* about its current state.[9]

The *a posteriori* probability distribution depends upon the actual state of the system and the rules that govern the interactions among its parts. It can be thought of as the probability that the system was, in the previous moment, in one of its possible states, given that we *do* know its current state.

Cause information is defined as the *distance* between these two probability distributions. There are various ways of measuring the distance between two probability distributions.[10] The distance measure should give zero for identical distributions and give greater values the greater the divergence in the two probability distributions. Intuitively, one can think of it as a measure of the work required to transform one distribution into the other. The cause information in a system is therefore a measure of the extent to which the system's current state constrains the probability distribution assigned to its past state space.

Thus, let $D(P_1 \mid\mid P_2)$ be the distance between probability distributions P_1 and P_2. Let $p(S^p)$ be the *a priori* probability distribution that assigns equal probabilities to all possible (past) states of system S. And let $p(S^p \mid S^c = s)$ be the *a posteriori* probability distribution for S's possible past states given its current state s. The cause information (ci) of S is then given by:

$$ci(S^p|S^c = s) = D(p(S^p|S^c = s)||p(S^p)). \tag{1}$$

We will return to this definition later, to consider what motivates its role in modeling consciousness (Sec. 3.1), and whether it can be considered a fundamental quantity (Sec. 4.1). The remaining notions of integration, and Q-shape, will now simply be defined intuitively.

A system's information is *integrated* if it is not determined by the information in its parts (treated independently). The relevant parts to consider are defined by the *minimum information partition* (MIP). This is the partition into independent parts that leaves the least information unaccounted for by the parts. The *amount* of integrated information in the system represents the amount of information in the system that is not determined by the information in its parts defined by the MIP. This amount is represented by ϕ.[11]

For a system S with $\phi = N$ ($N > 0$), it is possible that S has a part with $\phi > N$, and it is possible that S is part of a larger system with $\phi > N$. In either case, S's ϕ is not the theoretically interesting quantity. For such overlapping systems, the interesting quantity is the ϕ of the system with the largest ϕ, or ϕ^{max}. (Under Orthodox-IIT, only ϕ^{max} is a measure of consciousness.)

[8] The reason for taking the minimum is explained in Oizumi et al. (2014, pp.7-8). Note that information has broader application than is presented here. For example, we may speak of the information about a system contained in one of its subsystems. This is a measure of the extent to which the subsystem constrains the past and future states of the system. This will not be needed in what follows.

[9] One might think that there are other constraints (other than its present state) that suggest that the *a priori* distribution should be something other than an equiprobable one. Here it is important to note that the current mathematical formalism is not immutable, and revisable in light of such considerations.

[10] See e.g. Tononi (2012, pp.319-20, note 7).

[11] For a formal definition of ϕ see Oizumi et al. (2014, pp.8-13).

Finally, a *Q-shape* is specified by the informational relationships generated by the system (Balduzzi & Tononi (2009)). For a ϕ^{max} system, one can draw up its informational relationships in a space called "qualia-space" or "Q-space". The dimensions of Q-space correspond to the possible states of the system. The dimensions are of unit length, corresponding to probability one. A point in Q-space therefore picks out a probability distribution. Thus, one can identify many points in this space corresponding to how different parts of the ϕ^{max} system constrain the system's probability distribution. Together these points will create a shape, a Q-shape. (Under Orthodox-IIT, each Q-shape determines a specific qualitative character or quale.)

2.2 The empirical content

The predictions of Neutral-IIT all concern correlations between integrated information and the observable symptoms of consciousness. As subject's phenomenal reports are the paradigm of such observable symptoms, the predictions will be phrased in terms of them. There are two primary predictions:

> (i) Subjects' reports on when they are conscious are correlated with the presence of maximally integrated information (ϕ^{max}) states in those subjects.
>
> (ii) Subjects' reports on the qualitative character of their conscious experience are correlated with the Q-shape of their ϕ^{max} states.

It will also be useful in what follows to define an optional additional prediction:

> (iii) Subjects' reports on the extent to which they are conscious are correlated with the amount of information integration in those subjects.

Neutral-IIT deliberately imposes weak constraints on what counts as "correlated". One reason for this is that phenomenal reports must be treated with care, as they are not always reliable. But perhaps the main reason for the weak constraints is that the nature of the correlations will be clarified in different ways depending on one's interpretation. The predictions can become more constrained as the set of possible interpretations becomes more constrained by empirical, mathematical, and conceptual advances.

By focusing on (i) we can see that Neutral-IIT is falsifiable. Thus, if we find that the information in human brains invariably has zero integration, whenever such humans report having phenomenally rich conscious experiences, then Neutral-IIT is falsified. However, there are a number of situations left open to interpretation. Firstly, Neutral-IIT allows that conscious experience only arises when ϕ^{max} reaches some large threshold. Secondly, Neutral-IIT allows that artificial systems can be built to generate phenomenal reports (or analogues), despite having zero ϕ. Thirdly, Neutral-IIT allows that artificial systems can be built to generate large ϕ despite not being conscious, since it allows that consciousness arises out of information integration combined with some other (perhaps yet to be discovered) factors.

Now consider (ii). The basic idea is that the (reported) structure in subjects' phenomenology corresponds to (Q-shape) structure found in their ϕ^{max} states. But the correspondence need only be a mapping that enables the prediction of the reports given the Q-shapes. There need not be structural *similarity* between the (reported) phenomenal structure and the Q-shape structure.[12]

[12] One potential barrier to the philosophical neutrality of (ii) concerns how fine-grained Q-shapes are. In IIT it is nearly impossible for any two individuals, or any one individual at two distinct times, to have identical Q-shapes. This would seem to rule out the possibility of any two individuals, or any one individual at two distinct times, having identical qualia. If one's philosophical theory of consciousness entails that this is in fact possible,

Finally, prediction (iii) is optional. Reports on the extent to which we are conscious are rare, and open to interpretation. For example, we sometimes speak of consciousness diminishing as we fall asleep. But this provides little data, and could be understood in various ways. Indeed, Tononi (2008, p241) suspects that our ability to judge consciousness levels might be poor in an analogous way to our ability to judge temperature levels. We are good at judging temperature if it fluctuates around familiar levels, but not for levels outside that familiar range. We may therefore need a ϕ-measure for the same reason we need a thermometer. Indeed, if we have empirical justification for prediction (i), then since ϕ comes in degrees, we might thereby have abductive justification for consciousness coming in degrees, whether or not we can make good *a priori* sense of this claim. On the other hand, if one is simply opposed to talk of consciousness levels, one could simply drop (iii). For it is conceivable that future research could yield strong empirical support for (ii), yet only support (i) in the sense that consciousness emerges when ϕ^{max} reaches a certain threshold.

Even given Neutral-IIT's weak readings of (i)-(iii), the question remains: why would anyone take them seriously in the first place? After all, we simply do not have the technology to measure the ϕ (and hence, the Q-shapes) in any human brain components. Nor is such technology in the foreseeable future. So why even put time and money into developing IIT? To answer this question, we need epistemological interpretations.

3 Epistemological interpretations of IIT

3.1 Epistemology I: the orthodox axiomatic method

The starting point for Orthodox-IIT is encapsulated in the following statement: "As recognized by Descartes, my own experience is the only thing whose existence is immediately and absolutely evident" (Tononi et. al. 2016: p451). This statement contains both an ontological claim (consciousness exists, illusionism is false), and an epistemological claim (experience is known with a special kind of certainty).

According to Orthodox-IIT, IIT is justified in a two step process. In the first step, a set of five axioms are formulated. From the glossary of Oizumi et. al (2014, p4), axioms are defined as follows:

> *Axiom*: "Self-evident truth about consciousness. The only truths that, with Descartes, cannot be doubted and do not need proof."

Once these axioms are formulated a set of postulates about the physical mechanisms underlying consciousness are derived. The mathematical formalism is then based upon these postulates. From the same glossary, postulates are defined as follows:

> *Postulates*: "Assumptions, derived from axioms, about the physical substrates of consciousness, which can be formalized and form the basis of the mathematical framework of IIT."

Thus, the core evidence for IIT is derived from the purported self-evident nature of the axioms. As stated by Oizumi et al. (2014):

then one might reject (ii), making Neutral-IIT not completely neutral. Thanks to Tim Bayne for pointing this out. In response, one could say that the different fine-grained experiential qualities predicted by the fine-grained differences in the Q-shapes of two seemingly identical experiences is just not cognitively accessed. Alternatively, one could weaken (ii) by correlating qualia only with coarse-grained Q-shape structure.

> "IIT starts from the fundamental properties of the phenomenology of consciousness, which are identified as *axioms* of consciousness. Then IIT translates these axioms into *postulates*, which specify which conditions must be satisfied by physical mechanisms, such as neurons and their connections, to account for the phenomenology of consciousness. It must be emphasized that taking the phenomenology of consciousness as primary, and asking how it can be implemented by physical mechanisms, is the opposite of the approach usually taken in neuroscience: start from neural mechanisms in the brain, and ask under what conditions they give rise to consciousness, as assessed by behavioural reports.".

Let's look at how this works. The aim is to derive the form of the physical substrate of consciousness from axioms about consciousness. The first axiom is the *existence axiom*, which states that consciousness exists. The corresponding postulate, the *existence postulate*, states that mechanisms exist, where a mechanism is a physical system with causal power. If we assume the causal power is necessary for physical existence, then the existence postulate is derivable, at least from the conjunction of the existence axiom and the claim that if consciousness exists then it has a physical substrate.

The second axiom is the *composition axiom*, which states that consciousness is compositional (structured), such that each experience consists of multiple aspects in various combinations. The corresponding postulate, the *composition postulate*, states that the physical substrate must be compositional, that the relevant mechanisms are combined into higher order ones. The postulate plausibly does follow from the axiom (again, in conjunction with conditional claim that if consciousness exists then it has a physical substrate).

But how much more about the structure of the physical substrate of consciousness can we derive from phenomenology alone? As I will now argue, very little, and certainly not enough to suggest the formal definition of information.

Consider the third axiom, and its corresponding postulate. The third axiom is the information axiom:

> *Information axiom*: "Consciousness is informative: each experience differs in its particular way from other possible experiences. Thus, an experience of pure darkness is what it is by differing, in its particular way, from an immense number of other possible experiences. A small subset of these possible experiences include, for example, all the frames of all possible movies."

The idea appears to be that it is self-evident (at least on phenomenological reflection), that experiences are informative and that they inform by exclusion. For example, if I walk into a room and see a blue wall, the experience informs by ruling out an experience of a red wall, an orange wall (etc.) and the extent to which the experience informs corresponds to the extent to which the experience rules out other possible experiences that I could have had when looking at the wall.[13] What can we derive from this that is relevant to describing the physical substrate of such experiences? Here is the information postulate:

> *Information postulate*: "A mechanism can contribute to consciousness only if it specifies 'differences that make a difference' within a system. That is, a mechanism in a

[13]Tononi is not consistent with the formulation of this axiom. On some formulations, the claim that consciousness is *informative* is replaced with the claim that consciousness is *specific*, see Tononi (2017a). However, this makes the axiom vacuous, effectively equivalent to the logical truth that a given experience is what it is by differing from what it is not (Cf. Bayne (2018)). And since no non-tautologous conclusion can be validly drawn from a tautologous premise, there is no hope of deriving a postulate from this axiom.

> state generates information only if it constrains the states of a system that can be its possible causes and effects. The more selective the possible causes and effects, the higher the information."

Here the talk of a state of a system constraining the causes (immediate past) of the system should remind one of equation (1) above, the equation for the cause information (ci) in a system. Indeed, this postulate strongly suggests the formal definition of information. But is the information postulate even suggested by, let alone derivable from, the information axiom? I think not. To see why, compare the information postulate with the following, made-up postulate:

> *Made-up postulate*: A mechanism can contribute to consciousness only if it specifies 'differences that make a difference' within a system. That is, a mechanism in a state generates information only if it excludes other possible states that it could have been in. The more possible states it excludes, the higher the information.

I claim that if any physical postulate is derivable from the information axiom, it is the made-up postulate. The axiom tells us that an experience informs by excluding alternative experiences that could have been had instead. The amount of information corresponds to the amount excluded. I infer from this that the physical system whose state generates the relevant experience contains information in the sense of excluding other possible states that the system could have instead been in. The more possible states it excludes, the greater the information. Thus, take the experience of the wall's blue colour. The physical system that is the experience's physical substrate could have been reconfigured to have generated an experience of red, or an experience of green etc. The more configurations it excludes by being in its actual configuration, the more information it contains. Indeed, we can imagine building a formalism for this. The information in the relevant physical system is equal to $\mathrm{Log}_2(N)$ where N is the number of states it could have been in.

I am not claiming that the made-up postulate should play any role in neuroscience. Rather, its role is to help make apparent the invalid inference from the information axiom to the information postulate. The inference is invalid because the axiom says nothing about past or future, cause or effect. These notions are simply smuggled into the information postulate in an *ad hoc* manner. The made-up postulate, by removing such talk, stays truer to the axiom. But the made-up axiom fails to provide any useful description of the physical mechanisms underlying consciousness. Thus, while it may be possible to derive some physical postulates from the axioms, they can never have rich enough content to form the basis of a theory of the physical substrate of consciousness.[14]

3.2 Epistemology II: the natural kind approach

If the orthodox epistemology fails, then we are left with the problem of justifying the basic idea behind IIT: why think that information integration has anything at all to do with consciousness (or its observable symptoms)?

After criticizing the orthodox epistemology, Bayne (2018) suggests an alternative epistemology for IIT, which he calls the natural kind approach. According to this approach, consciousness is treated as a natural kind, which manifests itself via the observable symptoms of consciousness. One aims to find an underlying mechanism that accounts for those symptoms. Consciousness is then identified with the underlying mechanism. Crucially, the axioms still play a role in Bayne's account. The axioms are not (necessarily) considered to be axiomatic in the sense of being

[14]The same problem will apply to the fourth, integration postulate. The unity of consciousness does not in any way suggest the integration of information in the relevant sense of information. For additional criticism of the orthodox epistemology see Bayne (2018) and McQueen (forthcoming).

self-evident essential properties of consciousness. But they play a role by being a subset of the set of symptoms of consciousness that stand in need of explanation. Thus, one symptom of consciousness would be *being integrated* (or more cautiously, being judged to be integrated).

There are two problems with the natural kind approach. The first is that it is far from clear that the axioms should be included in the set of consciousness symptoms that stand in need of explanation, especially in light of Bayne's own trenchant criticism of them. Bayne argues, quite persuasively, that each of the axioms either is too vacuous to place any constraint on a theory of consciousness, or is too controversial to be considered an axiom. Either way, they do not have the status of data that must be explained by any theory of consciousness. Here we have considered the information axiom in particular. It is entirely unclear whether it is even true that consciousness informs through exclusion. But even if it does, it is not at all clear that this is a property of consciousness, perhaps it is just a property of our cognition, and how it happens to use conscious states to form beliefs.

The second problem with the natural kind approach, at least for present purposes, is that it is not sufficiently neutral. The natural kind approach assumes that consciousness exists (as a natural kind). But as has been stressed already, there is no reason why illusionists cannot help themselves to Neutral-IIT. For example, in the case of Illusionist-IIT (McQueen, forthcoming), Neutral-IIT is used to describe the kinds of states that introspection is prone to misrepresent, and to formulate and test correlations between integrated information and the production and expression of those misrepresentations. But if the axioms play no role at all in the justification of Neutral-IIT, then why should we take it seriously?

3.3 Epistemology III: minimal fallibilism

Here I offer a minimal fallibilist solution. Fallibilist, because it rejects the idea of infallible, self-evident knowledge, and instead treats scientific hypotheses as extremely tentative. Minimal, because it provides just enough justification for Neutral-IIT to be rationally adopted (and adapted) by various alternative ontological interpretations, even illusionism (Sec 4).

The solution is broken down into four steps. *Step one* begins with experimental justification for the importance of complexity or interconnectivity measures of consciousness. *Step two* then requires researchers to focus in on a specific complexity measure that can be developed and tested. Here, it simply does not matter how the researcher discovers their measure. The task is simply to get various measures on the table so that they can be experimentally scrutinized. One researcher might formulate what they take to be essential properties of consciousness and then try to derive their measure from those. Another researcher might find insight for the structure of a particular measure in a dream. This is hardly without precedent in the history of science. The structure of the Benzene ring, after all, is said to have come to Kekule von Stradonitz in a dream. Clearly, dreams do not justify theories. But they can play a heuristic role in the context of discovering a theory. So too for Tononi's axioms and his derivations from them. *Step three* involves extrapolating predictions from the chosen measure. *Step four* involves experimentally testing those predictions. Let us consider each step in more detail.

Step one: evidence for complexity measures. The first step in motivating the bare theory involves finding empirical support for complexity, or interconnectivity measures of consciousness. There is indeed such support. Compare the cerebellum to the cerebrum. Although the cerebellum has far more neurons than the cerebrum, the cerebrum is crucial to (reports of) consciousness whereas the cerebellum seems irrelevant. If neuron number is not relevant to (reports of) consciousness, then what is? Finding a crucial difference between the cerebrum and the cerebellum may answer the question. And indeed there is a striking difference in their interconnectivity: probe a region of cerebellum and it has little effect on other regions, but probe a region of cere-

brum and it is has significant effects on other regions of the cerebrum. This provides *prima facie* empirical support for the hypothesis that (reports of) consciousness has something to do with interconnectivity. And there is plenty more empirical evidence along these lines (see e.g. Koch (2018)).

Step two: hypothesizing specific complexity measures. The next step involves cutting down the set of all possible complexity measures so that one can focus one's research on some manageable subset. It matters little how a researcher does this. Here "axioms" may play a heuristic role in focusing a researchers attention on some possible subset. And this is one way that the ϕ^{max} measure can enter the picture. Crucially, the axioms are helping to *discover* as opposed to *justify* more specific measures. There is an analogy here with the discovery, and subsequent justification, of planetary orbit models. Consider the beginning of the scientific revolution. Natural philosophers knew that orbits were required to predict planetary observations. But which orbits? Creative thinking enabled natural philosophers to discover different hypotheses, including Ptolemy's complex geocentric epicycles, Copernicus' circular heliocentric orbits, and Kepler's heliocentric elliptical orbits. It is said that Kepler was motivated to develop the heliocentric view due to a mystical belief in the mathematical simplicity of universe. This helped him discover a specific theory of orbits, which would *later* be justified when sufficient technology (powerful telescopes) arose. Similarly, advocates of the orthodox interpretation have a (mystical?) belief that the physical substrate of consciousness can be derived from self-evident phenomenological considerations. Such considerations are playing the crucial role of discovery, but not necessarily justification, which often must come later. This gets us to the first prediction of Neutral-IIT, which relates (reports of) the presence of consciousness to information integration. The hypothesis is tentative, but worth working on and developing if we think some such complexity measure is justified by the evidence discussed in step one.

Step three: hypothesize correlations between features of the specific measure and (reported) features of consciousness. Just as astronomers might try to derive new observable predictions from a newly hypothesized astronomical model, so too can neuroscientists try to derive new observable predictions from a newly hypothesized complexity measure. For if the hypothesized complexity measure really does correspond to consciousness (at least in the sense of prediction (i)), then it is reasonable to suppose that more specific features of that measure correspond to more specific (reported) features of consciousness. This is how we get to predictions (ii) and (iii). Thus, if the hypothesized complexity measure is an unbounded ratio scale, then we might predict that insofar as we are capable of making judgments about levels of consciousness, they will correspond to this scale.

Step four: test the structural correlations and revise the measure accordingly. If there is a complexity measure that can yield bold predictions (like the correlations in (ii)), then there is an empirically motivated research program. The program involves testing those correlations. If those correlations are vindicated, then Neutral-IIT is more fully confirmed, and is on its way to being a mature scientific theory. If they are not vindicated, then the research program is not yet to be abandoned. Instead, the empirical findings should be used to revise the original complexity measure. If some sort of complexity is crucial to consciousness, then this is a natural scientific way of finding it. And indeed, this is what is happening right now, in attempts to find signatures of Q-shapes in subjects' brains to predict what they say about the qualitative character of their experiences.[15]

[15]For example, see Tsuchiya et. al. (2017) and Haun et. al. (2017).

4 Ontological interpretations of IIT

In the context of consciousness studies, an ontology must provide a clear account of the place of consciousness in nature. This means solving the hard problem of consciousness (Chalmers, 1995a). Thus, if consciousness is non-fundamental and arises out of more fundamental physical processes (*realist physicalism*), then the ontology should identify those processes and explain how consciousness arises from them, thereby solving the hard problem. If consciousness is fundamental (*Dualism* or *Russellian panpsychism*), then the ontology should identify the fundamental principles relating consciousness to fundamental physical properties. If consciousness is denied (*illusionism*), then the ontology must solve the illusion problem by explaining the physical mechanisms that create the illusion of consciousness.

4.1 Ontology I: The orthodox panpsychist ontology

The orthodox ontology is expressed in most detail in Tononi (2008).[16] One of its most crucial claims is the identity between consciousness and integrated information. But the identity takes a specific form: every quale is identical to the Q-shape corresponding to a ϕ^{max} state, and every Q-shape corresponding to a ϕ^{max} state is identical to some quale (Tononi, 2008, pp.224-32).

However, due to the abstract nature of Q-shapes, it is not clear how this provides an ontology, for what is the ontology of Q-shapes? After all, the space within which Q-shapes are defined is not the ordinary 3D space within which neurons reside. Q-shapes are only defined in the high-dimensional space whose dimensions correspond to points in the system's state space. A point in the high-dimensional space then corresponds to a probability distribution over the state space. This is perhaps why Tononi then proceeds, after explaining this identity claim, to give his "provisional manifesto", in which he explains "implications of the IIT for the place of experience in our view of the world" (2008, pp.232-40). The three ontologically relevant claims arising from this analysis are that consciousness is *pervasive*, *fundamental*, and *intrinsic*. I will take each of these in turn.

> *Consciousness is pervasive*: "IIT implies that many entities, as long as they include some functional mechanisms that can make choices between alternatives, have some degree of consciousness" (Tononi 2008: p236).

Integrated information seems easy to come by in nature. Isolated atoms and molecules will have it insofar as their internal states constrain their past and future states. Various parts of the cerebrum that are not excluded by the cerebrum's ϕ^{max} region will presumably also contain integrated information. So will regions of the cerebellum that have greater ϕ than the cerebellum itself. Moreover, it is possible to build simple devices with greater ϕ than the human cerebrum (Tononi, 2014). If there is consciousness where ϕ is maximized, then consciousness will be pervasive (though not necessarily everywhere). We thus have a form of panpsychism.

> *Consciousness is fundamental*: "it exists as a fundamental quantity—as fundamental as mass, charge, or energy" (Tononi 2008: p233).

Tononi treats consciousness as fundamental because he identifies consciousness with (Q-shapes of) integrated information and treats integrated information as fundamental. He compares the conception of the fundamental level of the universe in terms mass-energy distributions in spacetime, with a conception of the fundamental level of the universe in terms of integrated information distributed throughout spacetime. They are described as "equally valid". Indeed,

[16] See also Tononi (2017b).

Tononi even speculates that "entities with high ϕ exist in a stronger sense than entities of high mass."

> *Consciousness is intrinsic*: (i) "a complex generating integrated information is conscious in a certain way regardless of any extrinsic perspective" (Tononi 2008: p233); (ii) "According to IIT, being implies 'knowing' from the inside [...] Describing, instead, implies 'knowing' from the outside." (Tononi 2008: p234).

The intrinsicness claim is important if we are to locate Orthodox-IIT within existing ontologies in the philosophy of mind. However, Tononi's use of 'intrinsic' is obscure and appears ambiguous between (i) *objective* and (ii) *metaphysically intrinsic* (hence quotes (i) and (ii)). The claim that one's consciousness is objective (does not depend on others' perspectives) is not illuminating, since it is accepted as obvious by all realist ontologies. But Tononi also appears to use this notion in a stronger sense, as it plays a significant role in his solution to the hard problem.

To see this, consider how Tononi proposes to answer the "Mary the Neuroscientist" thought experiment (Jackson, 1982), which is routinely used to express the hard problem (e.g. Chalmers 1995b). Tononi concedes that although Mary (while in her black and white room) has all physical information about colour experience, she does not have all information about colour experience. But what is she missing if she can fully describe all Q-shapes corresponding to all possible colour experiences? "Obviously, although a full description can provide understanding of what experience is and how it can be generated, it cannot substitute for it: *being is not describing.*" (p234). Thus, when Mary leaves the room, she is no longer simply trying to describe colour experience from the extrinsic perspective, she enters a state which *is* colour experience, enabling her to know it from the intrinsic perspective.

This is effectively the solution offered by the version of pansychism known as Russellian panpsychism (Mørch (2018)). According to this view, physical descriptions only describe extrinsic properties of objects. They fail to describe their intrinsic natures. For example, to say that a particle has mass is to say something about what it is disposed to do when it encounters other entities: the particle is disposed to resist acceleration when it encounters an applied force. But what is it about the particle *intrinsically*, that gives it this disposition? Physics does not say, since physics is restricted to the extrinsic perspective. According to the Russellian panpsychist, the only truly intrinsic property we know of is consciousness, so it is concluded that all intrinsic natures (even for particle mass!) are consciousness-like (hence the panpsychism). Mary, insofar as she only possesses descriptions of colour experience Q-shapes in her black and white room, has an incomplete picture. For she has only the extrinsic skeletons of the experiences. What she is missing is their intrinsic natures, which physical descriptions cannot capture. The philosophy of mind that best captures Orthodox-IIT, then, is Russellian panpsychism.

Here I raise two objections. First, I object to the intrinsicness claim. In particular, I argue that integrated information, as defined by Tononi, is a paradigmatically *extrinsic* property; but then identifying consciousness with integrated information (or the Q-shapes it determines) renders consciousness extrinsic, in contradiction to the axiom.

Secondly, I object to the claim that consciousness is fundamental because integrated information is fundamental. In particular, I argue that for a quantity to be fundamental it needs to be well-defined at all scales, but integrated information fails this requirement. This latter objection will in turn undermine the pervasiveness claim.

Recall the definition of information from section 2.1. First one defines the state space of the target system, the set of all its possible states. But what defines this space? For a *closed* system, this is straightforward. A closed system is not interacting with any other system. So its state space is all configurations of the system allowed by physical laws. But IIT does not typically

deal with closed physical systems, it deals with open systems that are heavily interacting with their environments. How do you define the state space of an open system?

The usual expositions of IIT tend to ignore this question, since such expositions are restricted to simple systems of logic gates. For example, Oizumi et. al. (2014, p5, figure 1) consider a system of four logic gates, A, B, C, and D. Rather than calculating the information in the closed system ABCD, they calculate the information in the open subsystem ABC. D is then treated as a "background condition". D does not affect the size of the state space of ABC, which still has eight members corresponding to the eight possible assignments of on/off states. But D does affect the assignment of probabilities, because depending on its state, D can turn A on or off. Thus, one takes the state of D as "fixed". So if we consider what prior state of ABC could have led to the current state of ABC, it had better be consistent with the fixed state of D, which may constrain the probability assignments.[17]

Let us now imagine that the open subsystem ABC is the physical substrate of one's consciousness. According to IIT, it must be a maximum of ϕ, for example, it must be that $\phi(ABC) > \phi(ABCD)$. Is one's consciousness intrinsic to ABC's state? Clearly not, since if we vary the state of D (an entity extrinsic to ABC), then we vary the ϕ of ABC's state, not to mention its Q-shape. But then one's consciousness is not intrinsic to its physical substrate. IIT's treatment of background conditions makes clear that one's state of consciousness does not supervene on the state of its physical substrate. So then in what sense is consciousness intrinsic?

Now consider the claim that consciousness is fundamental. For a property to be fundamental, it must be well-defined at all scales. But many problem cases arise once we consider realistic systems. Consider a system where the relative positions and momenta of its parts matter, and so must enter into the definition of the system's state space. This means that each member of the state space assigns positions and momenta to each part of the system. If we are defining a fundamental quantity, we cannot be vague. Hence, each member of the state space must be specified in terms of absolutely precise positions and momenta for each part of the system. But how can we do that in light of Heisenberg's uncertainty principle? According to this principle, there is a fundamental limit to the precision with which certain pairs of physical properties, such as position and momentum, can be determined.

If one is to hold on to the idea that integrated information is fundamental, there seems to be no other option but to opt for some sort of quantum definition of integrated information. The possible states of a system would then no longer be characterized in terms of the exact positions and momenta of objects. Instead, they would be characterized in terms of the possible wave-functions of the system.[18] But on the face of it, this looks destined to fail. The wave-function of the system will describe the system at the fundamental microphysical scale. But IIT requires that we calculate ϕ at every scale to find the scale at which ϕ is maximized. That means comparing the ϕ calculated at the microphysical quantum scale with the ϕ calculated at the mesoscopic neuronal scale. The latter calculations land us back in the original problem. Perhaps this scaling aspect could be revised leaving us with a purely quantum notion of integrated information. But now the problem is that the quantum measure seems entirely divorced from the empirical data that is supposed to support IIT. For example, part of the empirical support for IIT is the fact that it is able to distinguish the kind of interconnectivity found in the cerebrum that is lacking in the cerebellum. But this interconnectivity concerns classical interactions happening at the neuronal scale.

This concern in turn affects the orthodox claims about consciousness being pervasive and intrinsic. It is difficult to evaluate the claim that integrated information (and hence, conscious-

[17] This is explained in the supporting information: Ouzumi et. al. (2014, Text S2, Supplementary methods).

[18] More precisely, since we are dealing with open systems, they would be characterized in terms of the possible reduced density matrices of the system.

ness) is pervasive if information is not well-defined for most realistic systems. Similarly, when the Russellian panpsychist states that consciousness constitutes the intrinsic nature of the extrinsic properties of physics, it is assumed that those properties are well-defined. For if they were not, then consciousness would not be well-defined. But if consciousness constitutes the intrinsic nature of integrated information, then this negative implication follows.[19]

In general, for integrated information to even be a candidate fundamental quantity, its formalization must transcend mere wire diagrams. Until then, the orthodox ontology fails to provide a defensible interpretation of IIT. Let us now consider alternative ontological interpretations that allow (in fact, require) integrated information to be non-fundamental.

4.2 Ontology II: The phenomenal concepts strategy

According to the phenomenal concepts strategy (PCS), consciousness exists, is non-fundamental, and is identical to some non-fundamental physical states (its neural correlates). The hard problem is an inevitable consequence of the special way in which we think about and conceptualize our own conscious states. In particular, we think about our own conscious states in terms of special phenomenal concepts. What makes these concepts special is that they are *inferentially isolated*.

To start with, consider a concept that is not inferentially isolated, the concept 'bachelor'. If one has a thought that incorporates this concept, say, the thought that Jones is a bachelor, then one can infer various other thoughts, that Jones is a man, that Jones is unmarried, etc. These inferences happen because the concept of a bachelor is inferentially connected to other concepts (man, unmarried, etc). According to the PCS, most concepts have rich inferential connections.

Phenomenal concepts are deemed special because they are *inferentially isolated*. This means that they do not support the kinds of inferences described above. We characterize phenomenal properties intrinsically, in terms of what they are like in and of themselves. This intrinsic characterization isolates phenomenal concepts from other concepts. From the fact that I have phenomenology with a character like "this", I cannot infer much at all.

The PCS now solves the hard problem as follows. Consciousness seems hard to explain because we want to infer phenomenology from physical descriptions of brain activity. But such inferences are prevented by the conceptual isolation of phenomenal concepts. Nonetheless, there is a purely physical account of conceptual isolation. Therefore, although we cannot deduce consciousness from physical descriptions, we can deduce that this deduction failure is inevitable! And this is enough for realist physicalism.

Consider Mary the neuroscientist. In the black and white room she does not have the concept of phenomenal redness. Since this concept is inferentially isolated, Mary cannot construct this concept from the ones she has. When Mary leaves the room, she looks at a rose and her brain enters a state which allows her to construct the concept of phenomenal redness. For she can now define this concept as the concept of *being in "this" (neural) state*. According to PCS, inferentially isolated phenomenal concepts are physical mechanisms in the brain. So PCS is a form of physicalism. There are different versions of the PCS that offer different theories of these mechanisms (Balog (2009)).

I will not try to evaluate the plausibility of the PCS here.[20] Instead I will consider a possible form for a PCS interpretation of IIT (PCS-IIT).

It seems PCS-IIT cannot adopt the orthodox axiomatic epistemology. Phenomenal concepts are not conceptually connected to the concepts expressed in the axioms (information, integration,

[19] Mørch (2018) has demonstrated an inconsistency between IIT and Russellian panpsychism, which she fixes by revising the IIT exclusion postulate. However, the above considerations still apply to the revised theory. For further considerations against the fundamentality of integrated information, see Barrett & Seth (2011) and Peressini (2013).

[20] Though see Chalmers (2007) for critique.

etc.). For this reason, the postulates cannot be inferred from phenomenology alone.

One might weaken the PCS slightly. Perhaps phenomenal concepts are not completely isolated, perhaps certain abstract properties of the physical substrate of consciousness bleeds through into phenomenal concepts. And perhaps this explains the pull of some of the axioms. However, in light of the problems raised for the orthodox epistemology, it seems that PCS-IIT need not go in this questionable direction. And anyway, this would be against the spirit of the approach. PCS-IIT can set aside the axioms and adopt the minimal fallibilist epistemology.

The PCS-IIT research program will try to use Neutral-IIT to help describe the physical mechanisms that underly phenomenal concepts. On the one hand, Neutral-IIT provides PCS with a framework for locating those mechanisms, since they seem to only act on ϕ^{max} states. On the other, Neutral-IIT may also help to explain why those states are the only ones that engender phenomenal concepts. In return, PCS can offer Neutral-IIT a better solution to the hard problem than is offered by Orthodox-IIT.

Finally, it is open to PCS-IIT to embrace pervasiveness aspect of Orthodox-IIT. But it is also open to PCS-IIT to maintain that consciousness requires additional factors that entail that consciousness only arises when ϕ^{max} reaches a certain threshold. Given the problems raised against the fundamentality of integrated information above, the latter is likely a better option. The challenge then, is to spell out the additional factors. Our next ontology provides some options.

4.3 Ontology III: Representationalism

According to representationalism, consciousness exists, is non-fundamental, and is identical to certain representational properties. More precisely, conscious experiences are physical states which have phenomenology, or phenomenal properties (qualia). A given phenomenal property (a given quale) is identical to a certain representational property, the property of representing a certain content. The idea is that all facts about representational properties can be physically described, thereby justifying physicalism. Representationalism comes in a wide variety of forms.[21] Here I will focus on one version of representationalism that has interesting connections to IIT.

Two decades after formulating the Mary the neuroscientist thought experiment, Frank Jackson had a change of heart. He now concludes that Mary *can* deduce all facts from within her black and white room, including all facts about redness phenomenology. Jackson (2003) defends this turnaround in terms of a version of representationalism. After explaining it, I will use it to formulate a representationalist-IIT.

Jackson's representationalism begins with the transparency of experience thesis. Whenever we try to describe our phenomenology, we end up just describing properties of the (putative) objects that our experiences represent. For example, if we try to describe the phenomenology of an experience of a blue bottle, we say that it has a "bluish" phenomenology, with a "cylindrical" character. Or take the phenomenology of a pain, being described as "intense and throbbing". These notions apply to what the experience represents, not the experience itself. According to the transparency thesis, what's happening here is that when we introspect our phenomenology in the hope of describing it, we in a sense "see right through it", and end up just describing properties that the represented objects (e.g. the blue bottle, one's stubbed toe) are represented as having. From this, Jackson concludes that a given phenomenal property is identical to a given representational property, where the representational property is the experience's property of representing the world to be thus and so.

[21] For the landscape of representationalist views, see Chalmers (2004) and references therein. I am here only considering versions of representationalism that (using Chalmers' terminology) satisfy both *reductive* representationalism and *narrow* representationalism.

Representationalism faces a serious challenge. The challenge is to say why some representations (like experiences) have phenomenology whereas other representations (like beliefs) do not. Thus, only experiences represent "in a phenomenal manner", and the challenge is to give an account of this phenomenal manner in physical terms. To solve this problem, Jackson specifies five physical features of phenomenal representations intended to capture the distinctive "phenomenal manner" in which they represent. Interestingly, there is significant crossover here with Tononi's axioms.

Jackson's five key features of phenomenal representations are: (i) *Richness*: phenomenal representations encode so much more information than other representations (like beliefs, or sentences). For example, one's visual field typically encodes any number of beliefs concerning specific locations of colours, shapes, etc. (ii) *Inextricability*: this is equivalent to Tononi's integration axiom, as Jackson says, you cannot "prise" the colour bit from the shape bit of a visual experience. (iii) *Immediacy*: experiences do not represent by creating a distinct mental state that does the representing for it, experiences represent directly and immediately. (iv) *Causal impact*: experiences represent the world as being the cause of those experiences, sound experiences represent sounds as *coming from* their location, etc. (v) *Functional role*: experiencing plays a distinctive functional role, which includes updating one's beliefs about the world in an ongoing way.

According to Jackson, Mary can come to know all facts about what it is like to experience red, from inside the black and white room. For these are all facts about representing surface properties of objects in a phenomenal manner. A content is represented in a phenomenal manner if the representation has features (i)-(v). When Mary leaves the room she will only acquire new skills (new "know-how"), since she will have the new ability to manipulate her new phenomenal representation in her cognition, for example, in imagination and memory. But she does not learn any new *facts*. All facts about the representation were available to her in the black and white room. All facts about experience are therefore deducible from purely physical facts, and physicalism is vindicated.

I will not here evaluate Jackson's representationalism.[22] Instead, I will explain how a representationalist-IIT might work in the context of Jackson's representationalism.

There is clearly some cross-over between Jackson's representationalism and the orthodox epistemology. Both appeal to five essential features of consciousness. However, they play very different roles. The five features specified by Orthodox-IIT are supposed to be features of experience itself. Those features are then used to derive the structure of their physical substrate. But Jackson's five features are supposed to be features of how experiences represent things to be (with the exception of the fifth). It is not clear that Tononi's project of specifying essential features of the experiences themselves is consistent with the transparency thesis that grounds Jackson's representationalism. Still, the difference is a subtle one.

IIT and representationalism could be reconciled in a number of ways. One question is whether representationalist-IIT should concede that consciousness exists, at least to some degree, wherever ϕ is maximized. This would entail that certain very simple systems (like grids of XOR gates) would represent the world in a phenomenal manner. But if we hold on to Jackson's five features, we can resist this implication. After all, only two of Jackson's features (richness, and inextricability) resemble Tononi's axioms. Jackson could complain that Tononi has provided necessary but insufficient conditions for consciousness. In addition to an experience maximizing ϕ, it must in addition represent with immediacy, with causal impact, and must also play a certain functional role with belief. The latter requires that a system can only be conscious (i.e. represent in a phenomenal manner) if it is capable of having beliefs (something that the grid seems incapable of). In that case, representationalist-IIT is not committed to the pervasiveness of consciousness em-

[22] See Alter (2006) for criticism.

braced by orthodox-IIT. For presumably only particularly large ϕ^{max} states will be accessible to belief. And if the only integrated states of interest are high-level neural states accessible to belief, then integrated information need not be considered fundamental either. Representationalist-IIT therefore avoids the problems that plague orthodox-IIT. The following research program is suggested: try to figure out what the physical substrate of experience must be like if it is to support experiences that not only satisfy richness and inextricability, but also the rest of Jackson's five features.

4.4 Ontology IV: Illusionism

According to illusionism, consciousness does not exist. The hard problem of consciousness is replaced by the illusion problem: explain the physical mechanisms that give rise to the illusion. For example, according to Frankish (2016), the mechanism of *introspection* systematically creates representations that represent experiences as having phenomenology when they in fact do not. These representations in turn cause us to make nonveridical phenomenal reports, and to feel puzzled about how phenomenology could arise from physical processes.

To determine whether illusionism is capable of offering an ontological interpretation of IIT, we can ask the following diagnostic question: if future experiments were to reveal strong support for the predicted correlation between qualitative character reports and Q-shapes (prediction (ii) from section 2.2), could illusionists incorporate the data?

It seems the answer is yes. A natural way that illusionists could incorporate this data is by treating high-ϕ states as the states that introspection responds to with phenomenological misrepresentations. The content of those misrepresentations vary in accord with variation in the Q-shapes of those high-ϕ states. The illusionist might think that introspection responds only to states that reach a specific threshold for high-ϕ. Alternatively, if future research reveals evidence of correlations between reported amounts of consciousness and amounts of integrated information (prediction (ii)), then the illusionist could let ϕ be a measure of the strength of the introspective illusion. This idea is developed in detail in McQueen (forthcoming), and so will not be explored further here.

Illusionist-IIT does not face the objections faced by the orthodox ontology. In particular, it does not require integrated information to be fundamental. In fact, it only requires that the mathematical formalism applies (approximately) to introspected states, like neural networks. It also does not entail that any simple systems are conscious. An isolated molecule with nonzero ϕ, for example, is not only unconscious, it differs significantly from humans in that it has no introspective mechanism that could create belief in consciousness.

5 Conclusion

IIT is typically expressed in terms of the orthodox interpretation. Both the epistemology and the ontology of this interpretation are problematic, and should be rejected. But we should not therefore reject IIT. Instead, we should distinguish Neutral-IIT from its orthodox interpretation and find better interpretations. Contemporary philosophy of mind provides many resources for this task, as illustrated in the three considered cases, the phenomenal concepts strategy, representationalism, and illusionism.[23]

[23]For helpful feedback I would like to thank Nao Tsuchiya, Leonardo Barbosa, Tim Bayne, Ole Koksvik, and Gabriel Rabin. This project was funded by the Monash University Network of Excellence for Complexity and Causation in the Conscious Brain.

References

Aaronson, S., A. Bouland, L. Chua, and G. Lowther (2013) ψ-epistemic theories: The role of symmetry, *Phys. Rev. A* 88, 032111.

Albert, D.Z. (1992) *Quantum Mechanics and Experience*, Cambridge University Press.

Alter, T. (2006) Does Representationalism Undermine the Knowledge Argument? In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Torin Alter & Sven Walter (eds.), Oxford University Press.

Balduzzi, D., & Tononi, G. (2009) Qualia: the geometry of integrated information, *PLoS Computational Biology*, 5(8), e1000462.

Balog, K. (2009). Phenomenal concepts, in the *Oxford Handbook of Philosophy of Mind*, McLaughlin, B.P., A. Beckermann, and S. Walter (eds.), Oxford University Press.

Barrett, A.B. & Seth, A.K. (2011) Practical Measures of Integrated Information for Time-Series Data, *PloS Computational Biology*, 7(1), e1001052.

Bayne, T. (2018) On the axiomatic foundations of the integrated information theory of consciousness, *Neuroscience of Consciousness*, 4(1): niy007.

Caves, C. M., C.A. Fuchs, and R. Schack. (2002) Quantum Probabilities as Bayesian Probabilities, *Phys. Rev. A* 65, 022305.

Chalmers, D.J. (1995a) Facing up to the Problem of Consciousness, *Journal of Consciousness Studies*, 2(3): 200-219.

Chalmers, D.J. (1995b) The Puzzle of Conscious Experience, *Scientific American*, 273(6): 80-6.

Chalmers, D.J. (2004) The representational character of experience, in Leiter, B. (ed.) *The Future for Philosophy.* Oxford University Press.

Chalmers, D.J. (2007) Phenomenal Concepts and the Explanatory Gap, in *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Torin Alter & Sven Walter (eds.), Oxford University Press.

Dennett, D.C. (1991) *Consciousness Explained*, New York: Little, Brown.

Dennett, D.C. (1998) Quining Qualia, in Marcel, A.J. & Bisiach, E. (eds.) *Consciousness in Modern Science*, pp.42-77, Oxford: Oxford University Press.

Dennett, D. (2003) Who's on first? Heterophenomenology explained, *Journal of Consciousness Studies*, 10(9–10), 19–30.

Dennett, D.C. (2016) Illusionism as the Obvious Default Theory of Consciousness, *Journal of Consciousness Studies*, 23(11-12): 65-72.

Frankish, K. (2016) Illusionism as a Theory of Consciousness, *Journal of Consciousness Studies*, 23(11-12): 11-39.

Griffiths, D.J. (2018) *Introduction to Quantum Mechanics, 3rd Edition*, Cambridge University Press.

Haun, A. M., Oizumi, M., Kovach, C. K., Kawasaki, H., Oya, H., Howard, M. A., Adolphs, R., Tsuchiya, N. (2017) Conscious Perception as Integrated Information Patterns in Human Electrocorticography, *eNeuro*, 4(5), ENEURO.0085-17.2017.

Jackson, F. (1982) Epiphenomenal Qualia, *Philosophical Quarterly* 32: 127–136.

Jackson, F. (2003) Mind and Illusion, in *Minds and Persons*, A. O'Hear (ed.), Cambridge University Press.

Koch, C. (2018) What Is Consciousness? *Scientific American* 318, 6, 60-64 (June 2018) doi:10.1038/scientificamerican0618-60.

McQueen, K.J. (2015) Four Tails Problems for Dynamical Collapse Theories, *Studies in History and Philosophy of Modern Physics* 49: 10-18.

McQueen, K.J. (forthcoming) Illusionist Integrated Information Theory, *Journal of Consciousness Studies*.

McQueen, K.J. and Vaidman, L. (forthcoming) In defence of the self-location uncertainty account of probability in the many-worlds interpretation, *Studies in History and Philosophy of Modern Physics*, ISSN 1355-2198, https://doi.org/10.1016/j.shpsb.2018.10.003.

Mørch, H.H. (2018) Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? *Erkenntnis*: https://doi.org/10.1007/s10670-018-9995-6.

Oizumi, M., Albantakis, L., & Tononi, G. (2014) From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLos Computational Biology*, 10(5), e1004654.

Peressini, A. F. (2013) Consciousness as Integrated Information: A Provisional Philosophical Critique, *Journal of Consciousness Studies*, 20(1): 180-206.

Nagel, T. (1974) What is it Like to be a Bat? *The Philosophical Review.* 83(4), 435-450.

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

Tononi, G. (2008) Consciousness as Integrated Information: A Provisional Manifesto, *The Biological Bulletin*, 215(3), 216-242.

Tononi G. (2012) Integrated information theory of consciousness: an updated account, *Arch. Ital. Biol.* 150, 290–326.

Tononi, G. (2014) Why Scott should stare at a blank wall and reconsider (or, the conscious grid). https://www.scottaaronson.com/tononi.docx. Accessed January 6 2018.

Tononi, G. (2017a) The integrated information theory of consciousness: an outline, in *The Blackwell companion to consciousness, 2nd edition*, Schneider, S. and Velmans, M. (eds.), John Wiley & Sons Ltd.

Tononi, G. (2017b) Integrated information theory of consciousness: some ontological considerations, in *The Blackwell companion to consciousness, 2nd edition*, Schneider, S. and Velmans, M. (eds.), John Wiley & Sons Ltd.

Tononi, G., & Koch, C. (2015) Consciousness: Here, There, and Everywhere? *Philosophical Transactions of the Royal Society of London*, Series B, Biological Sciences, 370(1668).

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016) Integrated Information Theory: From Consciousness to its Physical Substrate, *Nature Reviews, Neuroscience*, 17(7), 450-461.

Tsuchyia, N. (2017) "What is it Like to be a Bat?" - a Pathway to the Answer from the Integrated Information Theory, *Philosophy Compass*, 12:e12407.

Tsuchiya, N., Haun, A., Cohen, D., & Oizumi, M. (2017) Empirical Tests of Integrated Information Theory of Consciousness, in A. Hagg (Ed.), *Return of Consciousness.* Axon Foundation: Sweden.

Wallace, D. (forthcoming) What is Orthodox Quantum Mechanics? *The proceedings of the XII International Ontology Congress.* arXiv:1604.05973 [quant-ph]