Garrett Mindt

The Problem with the 'Information' in Integrated Information Theory

Abstract: *The integrated information theory (IIT) of consciousness is becoming an increasingly popular neuroscientific account of phenomenal experience. IIT claims that consciousness is integrated information in a system. I set this theory against the hard problem of consciousness (Chalmers, 1996; 1995) as the goal for a theory of consciousness to meet. In this paper I look to examine and ultimately critique IIT's use of the notion of information to base a theory of consciousness. I argue that the notion of information in IIT is a purely structural-dynamical notion, and so falls afoul of the structure and dynamics argument (Chalmers, 2003). I bolster these claims by appeal to the explanatory gap argument and show how IIT succumbs to this argument as well. For these reasons, I call into doubt IIT's ability to answer the hard problem of consciousness. Although this paper argues against the notion of information in IIT, in a broader context the criticisms which I raise here can be brought against any theory that attempts to explain consciousness as an information-theoretic phenomenon.*

Keywords: consciousness; integrated information theory; hard problem of consciousness; hard problem of information; structure and dynamics; explanatory gap.

Correspondence:
Email: Mindt_Garrett@phd.ceu.edu

Journal of Consciousness Studies, **24**, No. 7–8, 2017, pp. 130–54

1. Introduction

Ever since David Chalmers (1995) first introduced what he called the hard problem of consciousness it has been seen as a goal for a theory of consciousness to meet. The hard problem is the problem of explaining why there is any experience associated with all the physical processes going on inside our brains. There may be an elaborate story to explain *how* this might occur, such an explanation would consist of elaborating the structure and dynamics of that physical system (what Chalmers calls the easy problems), but such an explanation doesn't seem capable of answering the *why* question — why it feels like something for our brains to carry out all these physical processes (the hard problem). This paper will be looking at one attempt to explain the *how* and *why* questions of experience, *integrated information theory (IIT) of consciousness*.

In this paper I will be examining the foundations of IIT, specifically, how IIT defines and utilizes the notion of information as a base for a theory of consciousness. It has been argued (Chalmers, 2003; 1996) that physicalist accounts — those accounts which say the brain is wholly physical, and thus describable purely in terms of structural and dynamical features — are unable to offer a solution to the hard problem, since at most they will only ever explain more structure and dynamics, but fail to give an explanation of *why* there is any phenomenal experience associated with those physical processes. Through my discussion of IIT's use of information I will show that IIT is committed to a structural-dynamical (physicalist) notion of information and so falls victim to a number of anti-physicalist arguments.

I first introduce IIT in §2 and give a short account of the basic essence of the theory. Then I move on to give an overview of the account of information given within IIT, and suggest that in its current formulation it is a purely physical notion of information, and because of this IIT faces a number of problems commonly raised against physicalist accounts (§3). I argue that this account of information is solely structural and dynamical, and so has the same explanatory power as other physicalist accounts of consciousness. In the next section (§4) I elaborate on the explanatory gap argument and show how IIT succumbs to this argument as well. Therein, I also call into question some of the predictions (Tononi *et al.*, 2016) of IIT based on the aforementioned explanatory gap worry.

I conclude by rephrasing the hard problem of consciousness in terms of information. Since information-theoretic theories, such as IIT, claim that consciousness is the result of information (specifically how information integrates and is carried through a system), we might call the resulting problem for such theories *the hard problem of information: why is it the case that there is any experience associated with the informational processes occurring in our brain?* For information-theoretic accounts like IIT this is the heart of their hard problem, one must explain: (i) why particular organizations of information produce phenomenal experience in the brain, while other organizations of information, such as the laptop I am currently writing this essay on, produce none; and furthermore (ii) such explanations, to address the hard problem, must do so not merely through a purely structural-dynamical explanation. I do not think this is solely an issue for IIT, but rather for any account of consciousness that attempts to explain phenomenal experience as an information-theoretic phenomenon. If it can be shown that IIT's notion of information is insufficient to provide a foundation for a theory of consciousness, then IIT should revise the notion of information it utilizes in constructing the theory.

2. What is Integrated Information Theory?

IIT proposes that consciousness is integrated information in a system, the degree of which is signified by the Greek letter Φ.[1] The quantity of integrated information — or consciousness — present in a system is quantified by Φ which is 'the amount of information generated by a complex of elements, above and beyond the information generated by its parts' (Tononi, 2008, p. 216). The substantial difference between IIT and other philosophical or neuroscientific theories of consciousness is that it recognizes the significant amount of data given to us in our everyday experience. According to IIT we can use these data in constructing an account of consciousness, one that gives us a physically realizable model of consciousness. Having such a model would be a giant leap forward in our understanding of the mind, as it would give us the ability to quantify consciousness and so measure and study it scientifically. This would presumably lead to us having the ability to detect and predict when consciousness is present in a

[1] Tononi writes, 'Integrated information is indicated with the symbol Φ (the vertical 'I' stands for information, the circle 'O' for integration)' (Tononi, 2008, p. 220).

system (Tononi *et al.*, 2016). Aside from the ability to quantify the degree of consciousness present in a system, IIT might have interesting implications for certain empirical cases, such as split-brain cases (Tononi and Koch, 2015; Tononi *et al.*, 2016), and dissociative and conversion disorders (Oizumi, Albantakis and Tononi, 2014). The predicative and explanatory power of IIT gives one strong motivation to take IIT seriously, but only if integrated information is indeed identical to consciousness.[2] IIT makes the claim that consciousness can be captured in terms of varying quantities of integrated information, so we must be certain that the thing being quantified is indeed consciousness, and not merely integrated information itself. If one has good reason to think that when one quantifies integrated information one fails to quantify the degree of consciousness present in a system, then we have reason to suppose IIT is not a full explanation of consciousness.

IIT is constructed by first outlining what Tononi takes to be the five undeniable attributes of conscious experience (phenomenological axioms): *intrinsic existence*, *composition*, *information*, *integration*, and *exclusion*. According to IIT these axioms are evident to us through our experience, and so the theory takes them as axiomatic in constructing a theory of consciousness.[3] Tononi thinks that these phenomenological axioms are evidence enough to then derive a set of physical postulates which explain how these aspects of our phenomenology can be realized through a physical system, e.g. the brain. Since IIT is a neuroscientific theory its aim is to provide a detailed account of how consciousness is brought about by physical systems; it is the job of the physical systems postulates to give such an account. How might physical systems have the ability to bring about the essential aspects of our phenomenology (phenomenological axioms)? Presumably, according to IIT, this question is answered by the

[2] It is important to note here that Cerullo (2015) calls into question the explanatory power of IIT, regardless of its ability to tackle the hard problem of consciousness. Cerullo argues that IIT is really a theory of proto-consciousness, and so any explanations it might offer regarding consciousness are really explanations of proto-consciousness. According to Cerullo, this doesn't seem to provide us any answers to the so-called easy problems of consciousness (easy problems are things such as: attention, the directedness of behaviour, the correspondence between memory and cognition, etc.).

[3] That is not to say that these axioms are exhaustive — Tononi and colleagues admit that there may be more than the current five in IIT as it stands now. I will be taking these for granted as they are not the focus of this paper, but one could find disagreement in the axioms and postulates.

constraints detailed in the physical systems postulates. To give an example of how the axioms relate to the postulates let us take a look at the second axiom and postulate of IIT — *composition*:

> Consciousness is structured: each experience is composed of phenomenological distinctions, elementary or higher-order, which *exist* within it. (Tononi and Koch, 2015, p. 7)

This axiom is meant to express the essential property of our conscious experience, that there are many phenomenal aspects to our experience at any given time. For example, say you are sitting at your favourite local coffee shop. Within your experiential field is a white coffee cup in front of you with a latte steaming inside. Within that experience you have the phenomenal distinctions of white-cup, white, cup, in front of, table, steam, etc., all creating a *composition* of phenomenal distinctions. According to IIT, for physical systems to be able to instantiate this *composition*:

> The system must be structured: subsets of system elements (composed in various combinations) must have cause–effect power upon the system. (*ibid.*, p. 7)

Understood this way, the composition which is given to us in our everyday experience is the result of cause–effect powers of elements in a system, which are able to bring about change to one another and the system as a whole, thereby revealing phenomenal distinctions. By 'cause–effect power' Tononi means the way in which those various elements interact with other elements in the system, and so *causes* state changes to those elements and the system as a whole; and how other elements in turn bring about *effects* on a particular element, thus changing the overall state of the system. For an element to 'intrinsically exist', as Tononi puts it, an element must have cause–effect power upon itself, and must make a difference to the overall character of the state of the system as its states evolve and change over time.

One may disagree with the translation of these axioms into postulates, whether generally about the move from these axioms to postulates or the way in which they are translated, but I will set aside these disagreements to bring into focus the problem being discussed in this paper. For now, this will serve to give a general idea of how IIT is developed. IIT begins with *identifying* the *essential properties* of our experience — phenomenological axioms — and derives postulates that explain how physical systems might realize these axioms — physical systems postulates.

The five features of phenomenology and their corresponding postulates lead Tononi to posit a central identity of IIT — this will be of particular interest in §3 & §4 in examining IIT's use and definition of information:

> According to IIT there is an identity between phenomenological properties of experience and informational/causal properties of physical systems… The maximally irreducible conceptual structure (MICS) generated by a complex of elements is identical to its experience… An experience is thus an intrinsic property of a complex of mechanisms in a state. (Oizumi, Albantakis and Tononi, 2014, p. 3)

What exactly do Oizumi and colleagues mean by 'maximally irreducible conceptual structure'? According to IIT, the brain is composed of billions upon billions of elements (neurons/neuronal groups) and these elements take the place of information states — states which express some degree of information in their processing through the system as they fire, activating various regions of the brain. These elements do not exist distinct from one another. Rather, they form integrated complexes that express information greater than the information generated by those elements independently of each other. According to IIT they would be maximally irreducible, as separating any of those elements from one another would decrease the amount of information which it is able to express. In this sense the whole is greater than the sum of its parts. This is what IIT means by integrated information: information which through integration with other informative elements in the system achieves a state that expresses more information than those elements did independently from one another.

To summarize thus far, experience according to IIT is identical to a MICS, those conceptual structures are composed of integrated information states, *ipso facto* experience is identical to integrated information states. Given that this is what IIT is arguing, the case must be made that the integration of information can give one a thorough account of phenomenal experience. If IIT can make such a case it would need to propose a direct response to the why-question of experience: why is it that integrated information states have a phenomenal character associated with their instantiation?

The identity of experience with the MICS is of central importance as it is due to this, depending on what information is according to IIT, that the theory may face a number of objections commonly raised against physicalism. An important thing to keep in mind from the short explanation of IIT which I have given in this section is that IIT bases its theory of consciousness on the notion of information. We now need to make clear what exactly information is and how IIT defines and utilizes the notion in constructing a theory of consciousness.

3. What is Information?

It is by no means uncontroversial what exactly is meant by invoking the notion of ‘information’, since there is an ambiguity in what exactly one means by ‘information’. Does one mean the common-sense understanding of information as something which informs, and so gives one meaning or understanding, i.e. a semantic notion of information? Or do we understand information in terms of syntax, i.e. *how*, in the sense of what way, does information flow through a system, rather than the *meaning* of that information? Or do we mean some combination of the two? And, what exactly would such a combination look like? I suspect this ambiguity has something to do with a widely held assumption that things which can be said to contain information must have some sort of meaning associated with them.

For the purposes of my argument, however, we need to understand what Tononi means by ‘information’ as he defines it in explicating IIT. Accordingly, I argue that if Tononi’s use of information is as I have outlined it in the following subsection (§3.1) then his brand of IIT is committed to a physicalist position and so succumbs to the same problems as physicalist accounts more generally.

3.1. What is Information According to IIT?

Claude Shannon, arguably the father of modern information theory/ communication theory, in his paper ‘A Mathematical Theory of Communication’, is concerned with what he calls ‘the engineering problem’ in communication. This problem can be summed up as: how does a particular state of the system specify a particular message from the range of all possible messages expressible by that system? Since a system incapable of producing a vast array of possible messages to be transmitted would have very little use in communication, the system must be able to instantiate different possible messages. For example, when you type a message into your smartphone, it is able to transmit messages such as: ‘hey, what’s up?’, ‘what time for dinner?’, ‘should I bring wine?’, etc., and that is because: (i) it is a system that is able to transmit a vast array of possible messages; (ii) the system on the receiving end is one that is able to receive a vast array of possible messages which are sent to it; and (iii) these particular possible messages are unknown at the time of design and so must be able to discriminate between a large number of eventual possibilities. As Shannon says,

> …semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon, 1948, p. 379)

For instance, take a six-sided die (D6) as an example. This system is composed of six possible states (the possibilities ranging from 1–6), all of which convey a particular message to whoever rolls the die. That particular system has the possibility of communicating a number of states, and so can be said to instantiate information equal to $\log_2(6) = 2.59$ bits of information.[4] For systems with more possible states, when a particular state is achieved they convey more information expressible as bits, i.e. a twenty-sided die (D20) is $\log_2(20) = 4.32$ bits. Another way to understand what it means for something to convey more ‘bits’ of information is to say that information is the reduction of uncertainty, and the more uncertainty is reduced by the system the more information that system expresses. In the case of the D6 and D20 there is more uncertainty in the D20 system than in the D6, and so when one possible state is picked out of either system the system with more possible states has a higher degree of uncertainty reduced by instantiating a particular state of the system, i.e. 4.32 bits > 2.59 bits.

Before the die is rolled in either case the system is in a maximal state of uncertainty[5] as it’s jumping around in your hand; but once it is thrown and lands on any of the possible states 1–6 or 1–20, in either

[4] $\log_2$ expresses that there are two possible outcomes, in other words ‘is’ the case or ‘is not’ the case, ‘yes’ or ‘no’. In the case of the die there are six possible outcomes, and whatever number the die shows, say it lands on the 6 side, can be said to be a ‘yes’ response to 6 and a ‘no’ response to 1–5. Another way to think of it is that it expresses the unlikelihood of 6 being chosen out of all possible options. This is expressible in terms of bits of information, which, in this case, the physical system of the die produces 2.59 bits of information.

[5] To clarify the meaning of ‘maximal state of uncertainty’ — I mean maximal as in the maximum amount of *possible states*, not infinitely possible outcomes, since the die is not a system that can express an infinite amount of possible states but, as in the examples above, only 6 or 20, respectively.

case, the overall uncertainty of the system is reduced and that particular state (message) conveys information equal to 2.59 or 4.32 bits respectively. When uncertainty has been reduced it is the same as information being expressed by that system. And the more uncertainty which is reduced the more information that is expressed.

Tononi diverges from Shannon's definition of information when it comes to physical systems instantiating information, as his conception concerns information integration. Tononi says his definition is vastly different from how information is used in common language and communication theory (communication theory is in reference to Shannon's notion of information), and merely stays true to the etymology of the term 'information' (Tononi and Koch, 2015, p. 8). Rather, for a physical system to instantiate information, he thinks it must '*specify* a cause–effect structure that is *the particular way it is*: a specific set of specific cause–effect repertoires — thereby differing from other possible ones (differentiation)' (*ibid.*). Cause–effect repertoires are all the possible ways a particular element or set of elements can bring about changes to a system, or be affected by other elements, or sets of elements, in that system; thereby differentiating themselves causally from other elements that have an impact and can be affected by other elements. If elements have a cause and effect that is different from other elements' cause and effect on that system, then it can be said to have a cause–effect repertoire. With this in mind, Tononi proposes a modified version of information that has a causal notion built in:

> [I]nformation refers to how a system of mechanisms in a state, through its cause–effect power, specifies a form ('informs' a conceptual structure) in the space of possibilities. (*ibid.*)

Accordingly, information must be able to affect the system and in turn be affected by other elements in that system. Tononi adopts *differentiation* to articulate this, thus showing the way in which the elements in the system specify particular cause–effect structures differing from other elements in the system. This form of information is reminiscent of Gregory Bateson's definition of information from his *Steps to an Ecology of Mind*, in which he gives the following causal definition of information:

> [T]he world of form and communication invokes no things, forces, or impacts but only differences and ideas. (A difference which makes a difference is an idea. It is a 'bit,' a unit of information.) (Bateson, 1972, p. 276)

To unpack the quote, what Bateson means by 'differences that make a difference' is that some system can be said to convey information if it can bring about a change of state in another system. If some difference in the state of one system can bring about a difference in another system, then information has been conveyed. Understood this way, information is instantiated in a system when it is able to make a difference to the other system, and constrain the possible past and future states of that system — only differences which make a difference count as information. This definition of information gives one a causal notion of information, one that characterizes information in terms of how it flows through and brings about changes to other elements in the system. But what about this leads to the phenomenology? Tononi defines the phenomenological axiom of *information* as:

> 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*). (Tononi and Koch, 2015, p. 6, emphasis in original)

What is it about information integrating that gets one the phenomenal distinctions which we experience? Consider this example: what is the difference between my experience of the view from the Chain Bridge in Budapest, overlooking the Danube, and from Tower Bridge in London, overlooking the Thames? Aside from the obvious geographical difference between the two, they both afford a unique set of possible experiences. In one I have the *possibility* of seeing the London Shard, in the other I have the *possibility* of seeing the Hungarian Parliament. My neurophysiology has to be a system that can at any point in its operation discern the *differences* between these two vistas and any innumerable amount of other objects of experience. I experience the view from the bridge and discriminate in my environment a vast amount of small differences, which overall reduce the amount of uncertainty in my experiential field. Distilled to its core, according to IIT, phenomenology is a complex field of difference relations. This differentiation is thus accounted for by the differentiation of the internal elements from one another, according to IIT. Our experience of the world presents us with a large array of information, and if the system which produces consciousness is able to do that, it must in some sense be capable of accounting for the informational states instantiated in experience.[6]

The key thing to take from Tononi's use of information is the notion of '*differentiation*'. The question then becomes: is it possible to give an account of information as '*differentiation*' that can capture the phenomenal character of our experience of the world? The following subsection (§3.2) is an examination of this question.

3.2. The Problems with IIT's Use of Information

If Tononi thinks that the way elements in a system express information is by way of differentiation, then one must be certain that differentiation, as Tononi has defined it, really captures what we want in explaining consciousness. Consider the example above once again. If, according to IIT, my experience is composed of a highly organized collection of difference relations, and this is all done in my brain through integrated neurons and neuronal groups (those mechanisms which instantiate information states), where in this story of differentiation does the phenomenal character of experience come from? Understanding information states as differentiation gives one the difference relations which exist between various elements in a complex. In other words, one can track the change of the mechanisms in a global context of the system by tracking their differentiation from one another. This, presumably, will be instantiated by different neurons/neuronal groups firing in particular locations in our brain/ nervous system — firings in different spatial locations — and neurons/neuronal groups firing at different times — firings in different temporal locations. Differentiation gives one an effective way to understand the relationship between these various elements in the system, which stand in unique informational relationships to other elements in the system.

[6] Cerullo (2015) characterizes integrated information in what he calls the *principle of information exclusion*, which is that the 'level of consciousness is directly related to the amount of perceptual possibilities ruled out by the system' (*ibid.*, p. 3). The characterization of phenomenology being a 'complex field of difference relations' and Cerullo's principle above are subtly different. Cerullo's relies on the exclusion of 'perceptual possibilities', whereas I take it IIT is concerned with the internal differentiation of the mechanisms which compose the system from one another. In this sense, I take it that Cerullo's principle doesn't quite capture what is meant by IIT's notion of integrated information. It is about the differentiation of the elements themselves from one another, and that becomes reflected in our phenomenal experience, not merely what is excluded from our perceptual experience.

If we are to understand 'information' in terms of 'differences which make a difference' then these various informational elements standing in a unique set of spatial-temporal relationships will also have various causal relationships, i.e. how those elements affect and are affected by other elements in the system, bringing about a range of possible states of the system. If IIT is claiming we should understand information in this way, then one is given a structural[7] and dynamical[8] account of information.

Are structure and dynamics alone enough on which to construct a theory of consciousness? It has been argued by David Chalmers (2003) that structure and dynamics alone will not suffice in giving a satisfactory account of consciousness. The problem with a physicalist account of the world is that it solely relies on structure and dynamics to construct a theory of consciousness — presumably such an account would give a detailed description of how physical elements and their spatial-temporal organization, along with how those elements evolve dynamically through the system, cause other elements to change. Such accounts are only able to appeal to more structure and dynamics, essentially providing a detailed explanation of *how* consciousness comes to be, but failing to provide an equally thorough explanation of *why* consciousness comes about in the first place. However, there doesn't appear to be any good reason to think that truths about consciousness are fully captured through appeal to only structure and dynamics (*ibid.*, p. 120). This has become known as the structure and dynamics argument (*ibid.*), namely that structure and dynamics alone are not enough to account for consciousness.[9]

To put the structure and dynamics issue specifically in terms of IIT, why should it be the case that there is anything it is like associated with the relevant structural and dynamical properties of information presented by IIT? If IIT is to be considered a full-blooded account of consciousness it should be able to offer a response to this question.

In essence, there is a gap in explaining how integrated information states, which express difference relations, give rise to phenomenology. One should not just take for granted that 'differences which make a

[7] Spatial-temporal relationships between physically instantiated information states.

[8] Range of possible cause–effects on that system, i.e. the states evolve and change dynamically over time given the cause–effect relationships of other elements those information states stand in a relation to.

[9] For a thorough overview of the structure and dynamics argument, see Alter (2016).

difference' can account for our everyday experience. If IIT is making the claim that the structure and dynamics of integrated information states in a system can account for experience, then it would appear that IIT encounters a serious problem to which it must have a response. As David Chalmers has put it, in expressing the hard problem of consciousness:

> …the structure and dynamics of physical processes yield only more structure and dynamics, so structures and functions are all we can expect these processes to explain. The facts about experience cannot be an automatic consequence of any physical accounts, as it is conceptually coherent that any given process could exist without experience. Experience may *arise* from the physical, but it is not *entailed* by the physical. (Chalmers, 1995, p. 12)

To apply the above quote directly to IIT, experience may *arise* from integrated information states, but it is not *entailed* by them. IIT seems to make the argument that if experience arises from the structure and dynamics of integrated information states, then it is entailed by those integrated information states, and so posits an identity to explain that entailment. Yet, this move should give one pause — just because consciousness might arise from integrated information does not mean that it is identical to integrated information. To echo the concerns raised by Chalmers with regard to physicalist accounts of consciousness and apply them to IIT: experience may arise from integrated information states, but that does not necessarily mean experience is entailed by integrated information states.

For example, recall from §2 that IIT posits a central identity that experience is identical to the MICS. It may be the case that the MICS is a result of how physical elements in a system that express information are integrated, but it is another thing entirely for that MICS to be *identical* to experience. If one is convinced that structure and dynamics alone are not enough to explain consciousness, and that IIT's definition of information is a purely structural and dynamical one, then there cannot be an identity between experience and the MICS. This is because one is left with a gap from the structural and dynamical properties of integrated information to those properties of experience. IIT as it is currently explicated seems to skip a step in positing this identity. IIT has given us a detailed account of *how* experience might arise from integrated information, but has yet to provide a convincing reason to suppose that experience is *identical* to integrated information. This leaves open the question of *why* experience is the result of integrated information, and so leaves open the hard problem of consciousness.

If we accept the view of information that Tononi appears to be advocating in IIT — a modified and further developed form of Bateson's definition — then, because of its use of information, we are left with a dilemma as to how integrated information accounts for the hard problem, as it tells us nothing of the story of how one gets from structure and dynamics to our everyday experience.[10] I have argued that this is a consequence of IIT's use and definition of information and not at all in the spirit of the goal of IIT more generally — namely the goal of being a theory of consciousness that attempts to tackle the hard problem of consciousness (Tononi and Koch, 2015, p. 5). If IIT maintains a structural and dynamical notion of information, it doesn't appear likely that IIT will be able to account for the hard problem of consciousness. In §4 I bolster the structure and dynamics argument against IIT that I have made in this section by appeal to the explanatory gap argument.[11] Before I move on to §4, I first want to discuss some criticisms which have been raised by Cerullo (2011) and Searle (2013) against IIT's use of information to explain consciousness.

3.3. Understanding the Distinctions — Syntax vs. Semantics and Structure and Dynamics vs. Phenomenal

Much of the debate regarding notions of information have centred around the distinction between mathematical formulations of information, such as Shannon's notion of information (such notions we can

[10] In a recent blog post by Scott Aaronson (2014), in discussion with Giulio Tononi's reply to the post, David Chalmers and Scott Aaronson came to a consensus that IIT might offer a response to what they called the Pretty-Hard Problem (PHP). The PHP is the problem of picking out and predicting when consciousness is present in a system. Of course, this would mean it doesn't answer the traditional hard problem, but it would still put IIT a bar above other theories of consciousness, in so far as it would provide a powerful predictive tool in the scientific study of consciousness.

[11] It has been suggested that IIT might be interpreted as a kind of emergentism. This may help IIT avoid the charge of being a purely physicalist account, but at a prohibitively high cost. Most physicalist accounts would deny strong emergence, since strong emergence is arguably inconsistent with the causal closure of the physical domain (Kim, 2005). I take it this would be a less desirable position for a defender of IIT. In his blog, Peter Hankins (2014) suggests that one of the defenders of IIT, Christof Koch, should rather hold an emergentist IIT view than a panpsychist one (as Koch claims himself to be, Koch, 2012). Even if IIT were seen as an emergentist theory, then one trades avoiding my argument against IIT for a brute fact of nature, and still IIT would not be a robust explanation of consciousness in any useful sense.

refer to as syntactic notions), and semantic notions of information that attempt to understand how information acquires/expresses meaning. For the purposes of understanding the notion of information as it relates to consciousness, I find this distinction inadequate, as it fails to make clear what is important about information as it relates to consciousness. Rather, I have opted to frame the discussion in the previous sections in terms of structure and dynamics vs. phenomenal. I have done this for two reasons. Firstly, thinking of information in a purely mathematical/syntactic sense leaves out a vital notion of causation from understanding the dynamic quality of information that is required for understanding consciousness. I take it that the structural/dynamical features of a system can be quantified mathematically, and its syntactic structure mapped, but merely mapping out syntactic structure seems to leave out the meaning of the causal claims which are more naturally discussed with regard to structure and dynamics. Secondly, it's not clear that semantics fully captures what we mean when we want to understand the phenomenal aspect of information, since it is not at all certain that semantics is all there is to the phenomenal, thus leaving an important feature of what we are attempting to describe unrecognized. For example, it's not clear that the phenomenal experience of colours, shapes, etc. have any semantic features which are essential to their being experienced.

I think an important distinction to bring up is one IIT uses itself: IIT stresses that it is necessary to distinguish between *extrinsic* notions of information and *intrinsic* notions (Oizumi, Albantakis and Tononi, 2014, p. 6). Here I think a parallel can be drawn between the syntax vs. semantics and the structure and dynamics vs. phenomenal distinctions: *extrinsic information* is concerned with syntax and semantics — how information can be *quantified* from an extrinsic perspective and what that information *means* from an extrinsic perspective — versus *intrinsic information* which is concerned with structure/dynamics and the phenomenal — how information is *organized spatial-temporally* and *evolves dynamically* over time, and *what that is like* for the element in the system from the internal perspective. Clearly, the structure and dynamics of information alone would not be enough to capture the intrinsic perspective, since ultimately structure and dynamics can be quantified extrinsically. IIT seeks to explain how a system might gain an intrinsic perspective given a sufficient degree of integration, but it's not clear that as a result of sufficiently complex structural and dynamical properties of information an intrinsic perspective necessarily pops up. This is why the distinction is so important, and for the purposes of my argument so damning. Since IIT's causal notion of integrated information as *differentiation* is purely structural-dynamical, it fails to fully capture the *intrinsic* perspective, but merely quantifies the *extrinsic characteristics* of that physical system.

The best way to get to grips with engaging with the conception of information within IIT, whether to defend or critique it, is thus to adopt the right distinction. In the case of this paper, that is structure and dynamics vs. phenomenal, rather than the traditional syntax vs. semantics. To make the need to focus on the right distinction more apparent, I would now like to discuss two objections which have been raised against IIT for explaining consciousness in terms of information, each of which has taken a syntax vs. semantics approach to the debate. I endeavour to show how attacking IIT on the grounds of syntax vs. semantics fails to: (i) meet IIT on its own terms, and so fails to argue against IIT's causal notion of information; and (ii) further highlights the need to adopt the structure and dynamics vs. phenomenal distinction, over the classic syntax vs. semantics distinction, when discussing the relationship between information and consciousness.

Cerullo (2011) and Searle (2013) have raised worries for IIT with regard to using information to explain consciousness. Searle argues in his review of Christof Koch's (2012) book, *Confessions of a Romantic Reductionist*, that information cannot be used to explain consciousness, because information is an observer-dependent phenomenon, rather than an observer-independent phenomenon. Observer-independent phenomena would be things like electrons, rocks, galaxies, etc. — those things which exist that do not require an observer, and so would quite naturally exist despite humans observing them. This is in contrast to observer-dependent phenomena such as sonnets, novels, or papers on IIT, etc. that require an observer to realize their existence. Searle takes it that explaining consciousness in terms of an observer-dependent notion, such as information, would inevitably lead to such an explanation being circular in nature.

Ultimately though, this fails to take into account what IIT's project is attempting to do — it looks to describe the intrinsic features of a physical system, i.e. characterize information from an internal perspective. Searle's conception of information is an extrinsic one, since his objection concerns the dependence/independence of objects/ systems relative to an observer. Koch and Tononi respond to Searle's objection to their use of information to explain consciousness thusly:

> IIT introduces a novel, non-Shannonian notion of information — integrated information — which can be measured as 'differences that make a difference' to a system from its intrinsic perspective, not relative to an observer. Such a novel notion of information is necessary for quantifying and characterizing consciousness as it is generated by brains and perhaps, one day, by machines. (Koch and Tononi, 2013)

Koch and Tononi are correct to point out that criticisms on the grounds that information is an extrinsic phenomenon are only appropriately brought against Shannonian notions of information. Objections on these grounds fail to argue against IIT's causal notion of information, since objections regarding the observer-relevance of information are only concerned with extrinsic notions of information. Because of this, Searle's argument fails to argue against IIT's notion of information, and thus fails to bring into question IIT's account of consciousness on these grounds.

Cerullo (2011) criticizes IIT's notion of information for similar reasons as Searle. Cerullo argues that it is not clear how invoking the notion of information[12] in IIT should be at all useful for IIT in the way Tononi wants it to be. Cerullo argues that if IIT is going to be a contender to account for the challenges for a theory of consciousness outlined by Chalmers (1996; 1995) then it must meet the constraints of structural coherence[13] and organizational invariance.[14] Cerullo concludes that IIT fails to meet these two constraints, and thus integrated information does not do the job that Tononi suggests it does. As Cerullo says, 'A purely data-defined theory of information such as Shannon's lacks the ability to link information with the causal properties of the brain… Only by including syntactic, and most importantly semantic, concepts can a theory of information hope to model the causal properties of the brain' (2011, p. 58). If IIT had a purely Shannonian notion of information in the theory, I suspect Cerullo would be correct but, as I explained in §3.1 and §3.2, IIT has a

[12] Cerullo claims that IIT is employing a notion of information such as C.E. Shannon but, as was explained in §3.1, IIT does not hold a Shannonian notion of information, so this might be an uncharitable characterization of IIT's notion of information. Here I wish to point out that, although I also agree with Cerullo that there is an issue with IIT's notion of information, I disagree on what that notion of information is and why IIT's notion of information is unsuitable to base a theory of consciousness on.

[13] This constraint is meant to express that there is a correspondence between awareness and experience.

[14] This constraint is meant to express that systems with the same functional organization will have identical experience.

causal notion of information, one that is much more reminiscent of Gregory Bateson's (1972) notion. As a result of this, I take it that the spirit of Cerullo's critique of IIT is on the right path, in so far as it points out that IIT's notion of information is problematic, but ultimately the critique is unsuccessful because IIT does not have a Shannonian notion of information.

Whilst Cerullo's and Searle's criticisms are on the right lines, in so far as they point towards an issue with the use of information in IIT, their critiques overlook the non-Shannonian notion of information in the theory. By engaging with IIT's non-Shannonian notion of information, the arguments which I have advanced herein constitute an improvement on those of Cerullo and Searle. Thus, perhaps the most important reason to diverge from the syntax vs. semantics distinction is because of the notion of information at work in IIT. Syntax vs. semantics discussions are more applicable to Shannon's notion of information (an extrinsic notion of information) which IIT claims it does not employ, and I have attempted to show that in this section. Since IIT argues it has a causal notion of information, I have chosen to frame the issue in terms of structure and dynamics vs. phenomenal, which I think more accurately gets at the heart of the issue for IIT's notion of information (an intrinsic notion of information). The following section (§4) bolsters the structure and dynamics argument against IIT that I have made in §3.2 by appeal to the explanatory gap argument.

4. The Gap between the Physical and Phenomenal

The explanatory gap argument takes the form of highlighting the epistemic gap between physical facts and phenomenal facts — to put it another way, they try to show that knowledge of all the physical facts does not lead one to knowledge of facts about our phenomenology. Generally, once the epistemic gap has been secured, those arguing against physicalism then infer an ontological gap. I take it that even just securing an epistemic gap between IIT's notion of information and experience will be enough to show the seriousness of the problem for IIT. In particular, showing an epistemic gap between physical facts and phenomenal facts would be particularly detrimental to IIT given that Tononi begins with evidence from our experience (phenomenological axioms) and translates those into how physical systems could bring about said experience (physical systems postulates). As was explained in §2, IIT begins with the evidence from our own experience and uses that evidence to develop its 'phenomenological axioms', which it then uses to derive a set of corresponding physical postulates for how physical systems realize those phenomenological aspects of our experience. If there is an epistemic gap resulting from IIT's use of information (one of the axioms and postulates), then there may be good reason to doubt whether the other four axioms/postulates would hold as well, as these axioms and postulates are defined in terms of information.

4.1. The Explanatory Gap Argument against IIT

The explanatory gap argument goes as follows:

1) Physical accounts explain at most structure and function.
2) Explaining structure and function does not suffice to explain consciousness.

---

3) No physical account can explain consciousness.[15]

Any physical account will involve an explanation of consciousness in terms of structure and functions because that is the purview of the physical sciences and, according to physicalism, all facts about consciousness are accounted for by physical facts. There are of course certain things which can have a full explanation in terms of structure and function, such as the fact that water is H_2O. Presumably an explanation of water as H_2O in terms solely of structure and function would be an exhaustive explanation of water. Such an explanation would be satisfactory since it tells us exactly why every instance of water is H_2O, and conversely why every instance of H_2O is water. Furthermore, such explanations of water and H_2O will also tell us at what temperature water/H_2O reaches a boiling point, at which point it freezes, what particular conditions must obtain for it to go through state changes, e.g. from a solid to a liquid, etc. None of these explanations require, nor hint towards, a grander explanation than the purely structural and functional one provided to us.

The problem with consciousness is that it doesn't seem to be the case that such an explanation purely in terms of structure and function would give us such an analogously exhaustive explanation. Recall that

[15] The argument, as it is formulated here, comes from Chalmers (2003); the original argument is given by Levine (1983).

IIT posits an identity between phenomenal experience and the ‘informational/causal properties of physical systems’ — the MICS. If one is to get an exhaustive explanation of consciousness in these terms, it should be analogous to the case of water being H_2O — one should be satisfied with the explanation that experience is information/causal properties, and conversely that information/causal properties are identical to experience.

If we are to understand function as ‘causal roles in the production of a system’s behaviour’, as Chalmers suggests, then I take it that ‘intrinsic cause–effect structures of certain mechanisms in a state’ (as explained by IIT) satisfy the relevant causal role in producing the behaviour of a system, as it is the intrinsic cause–effect structures that constrain the possible states mechanisms within which a system can instantiate, i.e. neurons in the brain. Furthermore, if we are to understand structure as spatio-temporal structures, then the overall ‘space of possibilities in their past and future’ — all those possible mechanisms arranged spatially and temporally (neurons to other neurons) — is the entire structure of the overall system. In its entirety this definition of what information is, and thus what consciousness is, consists in a specification of the structural and functional properties of a system. The structural and functional properties of a system are not enough to explain consciousness. If information according to IIT is about how a mechanism through its ‘cause–effect power’ and ‘space of possibilities’ is nothing over and above structure and function, then IIT is committed to being a physicalist account of consciousness. If this is so, IIT succumbs to the same explanatory gap argument against physicalism. The existence of an epistemic gap due to IIT’s use of a physicalist construal of information is an undesirable consequence to say the least.

Now let us give a revised explanatory gap argument specifically for IIT:

1) Integrated information theory explains at most structure and function.
2) Explaining structure and function does not suffice to explain consciousness.

---

3) Integrated information theory cannot explain consciousness.

One might object that IIT is not solely a theory based on its construal of information, it is just attempting to make sense of our phenomenology and apply that to how physical systems might instantiate phenomenal experience. So it may be objected that the theory is not lead by its construal of information, but rather the character of our own experience. The issue with this response is that IIT posits an identity between one's integrated information structures and conscious experience — which means it should cut both ways. Tononi's view is set up by taking as evidence our phenomenology and then positing physical systems postulates that are able to realize those phenomenological axioms in a physical system. But if one is not able to go the other way — start with the physical systems postulates and derive the phenomenological aspects of experience — then something is terribly amiss. If there is an *identity* between the integrated information states and phenomenal experience, there should be no gap whatsoever. I fail to see how the austere physical language used to describe the physical postulates leads one naturally to the phenomenological aspects of our experience.

To motivate the explanatory gap, and to further call into doubt the explanatory power of IIT if the argument I have just proposed holds true, I would now like to turn to one of the possible predictions IIT argues is a consequence of the theory and show why it might not actually have this predictive power. In a recent paper by Tononi and colleagues (Tononi *et al.*, 2016) they have argued that IIT offers explanations and predictions regarding the physical substrate of consciousness. Specifically, I am interested in one particular prediction they argue IIT makes regarding consciousness and its physical substrate: that 'consciousness should split if a single major complex splits into two or more complexes' (*ibid.*, p. 10). Let us grant that, because of how information integrates in the brain, the two hemispheres achieve a global maximum of Φ, and that when there is a bi-section of the corpus callosum this global maximum is separated into two distinct complexes. Despite this, there would still be an explanatory gap.

Such a prediction would seem to lend support to IIT as solving one of the so-called 'easy problems' — as mentioned previously, the easy problems of consciousness are those such as the directedness of behaviour, the relationship between language and thought and, more importantly, the integration of information in the brain (Chalmers, 1995). IIT give us an explanation of how information integrates in the brain, and the fact that it explains and predicates the result of split-brain cases seems to provide strong support for that. Yet, it doesn't tell us why there is anything it is like associated with that information integration. Such an explanation/prediction of the theory still doesn't bridge that gap. It tells us *how* information integration occurs across the two hemispheres, but not *why* there is anything it is like associated with that information integration.

Furthermore, for the sake of argument let us say that this prediction of IIT is tested empirically, and we find that when a major complex of integrated information splits into two separate complexes consciousness splits as well. This is essentially what occurs in split-brain cases, when there is a bi-section of the corpus callosum, leaving the two hemispheres of the brain detached from one another. Let's assume that IIT runs the experiments and confirms this prediction on behalf of IIT, and finds that when one major complex is separated into two complexes one has two local maximums of Φ and thus a separation of consciousness. Does IIT really provide an explanation of this? Cerullo (2015, p. 5) calls into doubt the explanatory power of IIT in this regard, by showing that his own *faux* theory of consciousness would have the same prediction as IIT. Cerullo proposes a *faux* theory which he calls Circular Coordinated Message Theory (CCMT). Cerullo says '[t]he justification for CCMT is the self-evident property that consciousness is related to information traveling in feedback loops within a system (the principle of information circulation)' (*ibid.*, p. 5), the value of the degree of information circulation is signified by Omicron (O). Both have the same prediction, that when a major complex of Φ or O is separated there will be two complexes each with a local maximum of Φ or O, respectively. Both seem to have equal explanatory power — one says that this can be explained because there is a great deal of information integration between the two hemispheres, the other because there are significant cortico-thalamic loops. Which explanation is better? It seems both IIT and CCMT have equal predictive power. This would seem to at least call into question the weight behind such predictions of IIT.

Falling victim to the explanatory gap argument is a serious shortfall of IIT, as the theory is constructed with the hard problem in mind as the target. I don't think this is solely an issue with the theory, but rather with the definition of information utilized by the theory, because this is what commits IIT to giving a purely structural and dynamical explanation of consciousness. If one could change the definition of information according to IIT, that could avoid these obstacles, then IIT would be in a more robust position to tackle the hard problem.

If IIT falls so easily into a gap because of how information is defined according to the theory, the whole theory shouldn't be scrapped, but rather the definition of information. I leave open what such an account of information might be, as fully developing and defending such an account is outside the scope of this paper.[16] The goal has been to merely highlight the issue in IIT's definition and use of information. The next step for IIT, now that such worries have been raised, is either to show that the arguments I have given do not hold, or take on board the worries raised and offer a revised notion of information. I see no reason a more amenable notion of information cannot be developed. Such a notion of information will put IIT on a better track to solve the problem it intends to account for — the hard problem of consciousness.

5. Conclusion

In this paper I have shown that IIT is committed to a purely structural/ dynamical notion of information, and because of this commits itself to a physicalist account of consciousness, thus leading IIT into a number of objections commonly brought against physicalist accounts. If IIT wishes to avoid these issues, which I argued there is good reason to think it should in §3 and §4, then it will need to rethink how it goes about defining information. The issues at play for IIT's definition of information are analogous to the issues at play in the hard problem of consciousness. One can capture the issues raised in this discussion of IIT's use of information as the *hard problem of information*: *why is it the case that there is any experience associated with the informational processes occurring in our brain?* The burden of proof falls to IIT and other information-based accounts of consciousness if they wish to avoid the issues raised in this essay.

I have endeavoured to show that IIT should look at these issues in defining information when using it to construct a theory of conscious-

[16] To at least indicate some possible notions of information that might be developed further to avoid these issues, one might look at Chalmers' (1996) dual-aspect account of information. Though it's not clear that a direct application of a dual-aspect account will avoid the worries raised in this paper, there may be some promising developments that could come from exploring a dual-aspect account as it relates to IIT. Another option, and one which Cerullo (2011) discusses, is the General Definition of Information (GDI) from Floridi (2009), though Cerullo dismisses it for the reason that it will face standard philosophical worries concerning meaning (Cerullo, 2011, p. 57). I share these concerns with Cerullo, as it doesn't appear GDI will be able to bridge the traditional syntax vs. semantics gap, nor the structure and dynamics vs. phenomenal gap discussed in this paper, but nonetheless there may be some interesting developments that can come from further looking into the GDI as it relates to consciousness.

ness. IIT would offer an incredible degree of explanatory and predictive power when it comes to consciousness if integrated information is in fact quantifying consciousness. If IIT can come up with an alternative notion of information, then perhaps it may one day account for the hard problem of consciousness.

Acknowledgments

I would like to take this opportunity to thank those that through discussion and comments have helped improve the ideas within this paper. I would especially like to thank Matthew Baxendale, Philip Goff, and Michele Luchetti for helpful advice, discussion, and thoughtful comments throughout the writing of this paper. I would also like to thank Larissa Albantakis, Heather Mackenzie Morris, Gregory Miller, and Grace Whistler for comments on various drafts. An earlier version of this paper was presented at the Neurophilosophy Group of the University of Milan and at SOPhiA 2016 at the University of Salzburg. I would like to thank the audiences at both occasions for their insightful questions and objections. Many thanks to two anonymous reviewers for helping to hone the paper, offering useful suggestions and pointing out areas for improvement.

References

Aaronson, S. (2014) Why I am not an integrated information theorist (or, the unconscious expander), *Shtetl-Optim.*, [Online], http://www.scottaaronson.com/blog/?p=1799 [1 Nov 2016].

Alter, T. (2016) The structure and dynamics argument against materialism, *Noûs*, **50**, pp. 794–815.

Bateson, G. (1972) *Steps to an Ecology of Mind*, Chicago, IL: University of Chicago Press.

Cerullo, M. (2011) Integrated information theory: A promising but ultimately incomplete theory of consciousness, *Journal of Consciousness Studies*, **18** (11–12), pp. 45–58.

Cerullo, M. (2015) The problem with phi: A critique of integrated information theory, *PLoS Computational Biology*, **11**, e1004286.

Chalmers, D.J. (1995) Facing up to the problem of consciousness, *Journal of Consciousness Studies*, **2** (3), pp. 200–219.

Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, Philosophy of Mind Series, New York: Oxford University Press.

Chalmers, D.J. (2003) Consciousness and its place in nature, in Stich, S.P. & Warfield, T.A. (eds.) *Blackwell Guide to the Philosophy of Mind*, pp. 102–142, Oxford: Blackwell.

Floridi, L. (2009) Philosophical conceptions of information, in Sommaruga, G. (ed.) *Formal Theories of Information: From Shannon to Semantic Information*

Theory and General Concepts of Information, Lecture Notes in Computer Science, pp. 13–53, Berlin: Springer.

Hankins, P. (2014) Not a panpsychist but an emergentist?, *Conscious Entities*, [Online], http://www.consciousentities.com/2014/01/not-a-panpsychist-but-an-emergentist/ [1 Nov 2016].

Kim, J. (2005) *Physicalism, Or Something Near Enough*, Princeton, NJ: Princeton University Press.

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

Koch, C. & Tononi, G. (2013) Can a photodiode be conscious?, *New York Review of Books*, [Online], http://www.nybooks.com/articles/2013/03/07/can-photodiode-be-conscious/ [28 Nov 2016].

Levine, J. (1983) Materialism and qualia: The explanatory gap, *Pacific Philosophy Quarterly*, **64**, pp. 354–361.

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

Searle, J.R. (2013) Can information theory explain consciousness?, *New York Review of Books*, [Online], http://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/ [5 Oct 2016].

Shannon, C.E. (1948) A mathematical theory of communication, *Bell System Technical Journal*, **27**, pp. 379–423.

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

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

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

Paper received July 2016; revised December 2016.