1-1-2019

# Illusionist Integrated Information Theory

Kelvin J. McQueen

# Illusionist Integrated Information Theory*

Kelvin J. McQueen†

July 5, 2018

**Abstract**

The integrated information theory (IIT) is a promising theory of consciousness. However, there are several problems with IIT's axioms and postulates. Moreover, IIT entails that some two-dimensional grids of identical logic gates have more consciousness than humans. Many have found this prediction to be implausible, and as will be argued here, this prediction also exacerbates the so-called "hard problem of consciousness". Recently, it has been argued that if we treat the phenomenological aspects of consciousness as an illusion (illusionism), we can avoid the hard problem altogether by replacing it with the more tractable illusion problem: the problem of explaining how introspection systematically misrepresents experiences as having phenomenology. IIT is intended to be a theory of the phenomenological aspects of consciousness. However, it is possible to reformulate the axioms and postulates of IIT consistently with illusionism. Here it is argued that the resulting theory - illusionist integrated information theory - removes several problems for IIT including the hard problem and the logic gate problem, and also enables meaningful progress for illusionists on solving the illusion problem.

## Contents

**1 Introduction** **2**

**2 Integrated information theory** **3**

**3 The hard problem of consciousness** **6**

**4 From the hard problem to the illusion problem** **8**

**5 Illusionist Integrated Information Theory** **11**

**6 Problem solving with Illusionist-IIT** **13**

**7 Towards a solution to the illusion problem** **15**

**8 Conclusion** **17**

*This is a preprint. The final and definitive version of this paper will be published in the Journal of Consciousness Studies.

†Department of Philosophy, Chapman University, Orange County, CA, USA.

# 1 Introduction

Neurophysiological and computational modeling provides excellent tools for specifying mechanisms that perform cognitive functions. These models can potentially explain cognitive functions relevant to consciousness, such as awareness, memory, reportability, and so on. However, Chalmers (1995, 1996) has argued that these tools are insufficient to explain the phenomenological aspects of consciousness. He argues that even if all cognitive functions relevant to consciousness are explained, we would still be left with the question of why the performance of these functions is accompanied by phenomenology. These phenomenological aspects of consciousness are not defined in terms of cognitive functions, but instead in terms of what it is like to experience them. A state is said to exhibit *phenomenal consciousness* if and only if there is something it is like to be in that state (Nagel, 1974). For Chalmers, the problem of finding mechanistic explanations of cognitive functions are the "easy" problems of consciousness, while the problem of explaining phenomenal consciousness is the "hard" problem of consciousness. The hard problem poses a serious problem for the development of a complete theory of consciousness. It also poses a deep philosophical puzzle concerning the relationship between our minds and the physical universe.

The integrated information theory of consciousness (IIT) was formulated by Tononi (2004, 2008) and has been advanced by a number of researchers over the last decade. IIT is explicitly a theory of phenomenal consciousness, but also provides an operational definition of phenomenal consciousness: it is that property that you lose when you fall into dreamless sleep and regain when you either start dreaming or wake up (Tononi et. al. 2016: p450). This is intended to include a variety of experiences including sense perceptions, moods and emotions, dream states, mental images, and so on.

In terms of precision, IIT is revolutionary. It is the first theory of consciousness to offer a measure of consciousness that is both empirically defensible and potentially applicable to any possible physical system.[1] IIT defines $\Phi$, which is a measure of the amount of integrated information in a physical system. Integrated information is a certain type of interconnectivity among entities (e.g. neurons) that is apparently present in conscious regions of the brain but absent in unconscious regions.

However, despite the outstanding precision of this theory and its burgeoning empirical applications (e.g. Tsuchiya et. al. (2017) and Haun et. al. (2016)), IIT faces a number of problems. Firstly, there are problems with IIT's axioms and postulates, to be explained in the next section. Secondly, IIT entails that a simple 2D grid of identical logic gates can have greater consciousness than a human (Tononi, 2004). Many find this prediction implausible. Thirdly, IIT does not make progress on the hard problem. It may well be that integrated information is the correlate of phenomenal consciousness in the sense that whenever there is phenomenal consciousness there is integrated information. But a correlation is not an explanation, it is something that stands in need of explanation. Why is integrated information accompanied by phenomenal consciousness? While the hard problem is an apparent problem for any neuroscientific theory of consciousness, it is argued in section 3 that IIT's prediction of conscious 2D grids exacerbates the hard problem by blocking standard responses to it.

Recently, it has been argued by Frankish (2016a, 2016b) that if we treat phenomenal consciousness as an illusion, then we eliminate the question of why any system would be accompanied by phenomenal consciousness, thereby removing the hard problem. This still leaves a residual problem behind, the problem of explaining how introspection systematically misrepresents experiences as having phenomenology. Frankish calls this the *illusion problem* and argues that it is much more tractable than the hard problem.

On the face of it, IIT is a theory of phenomenal consciousness that is deeply at odds with

[1] Fully generalized integrated information measures are still being developed, e.g. Tegmark (2016).

illusionism. However, it is possible to reformulate the axioms and postulates of IIT, consistently with illusionism. I call this reformulation *illusionist integrated information theory*. I argue that illusionist-IIT removes several problems that plague IIT, including the hard problem and the logic gate problem.

In addition, I argue that illusionist-IIT helps solve problems for illusionism too. Firstly, the empirical support for IIT is growing. For example, it has recently been claimed that it is possible to predict a subject's experience just by considering the patterns of integrated information in the subject's brain.[2] If there is an empirical case for integrated information being the physical correlate of phenomenal consciousness, then there is a corresponding puzzle for illusionism: if phenomenal consciousness does not exist, then why does it appear to have a stable physical correlate in the brain? As we shall see, illusionist-IIT has a built in solution to this puzzle.

Secondly, I will argue that illusionist-IIT enables progress on the illusion problem. Frankish (2016) tries to explain the illusion by appeal to *phenomenal concepts*. The idea is that introspection systematically represents experiences in terms of defective phenomenal concepts. However, it is unclear what the content of these concepts are, and how we could have acquired them. It is argued that we can better understand phenomenal concepts if we treat them as partially veridical such that their veridical content is integrated information in the brain. In section 7 I argue that this enables progress on the illusion problem by making it easier to see why the illusion problem is an "easy" problem of consciousness and by creating avenues for empirical research on the mechanisms underlying the introspective illusion.

In the next section I explain IIT. Section 3 then explains the hard problem, the logic gate problem, and why the latter problem exacerbates the former. Section 4 introduces illusionism and explains how it replaces the hard problem with the illusion problem. Section 5 reformulates IIT into illusionist-IIT. Section 6 formulates seven distinct problems for IIT and argues that illusionist-IIT resolves them. Finally, section 7 argues that illusionist-IIT makes progress on solving the illusion problem.

## 2 Integrated information theory

IIT assumes that introspection gives accurate knowledge of phenomenology[3]. In fact introspection is regarded as so accurate, that IIT relies on it to formulate the basic axioms from which IIT's physical postulates are derived. IIT has five axioms. These axioms are intended to be self-evident facts about consciousness that can be verified simply by introspecting one's own conscious experiences. In what follows consciousness always means phenomenal consciousness.[4]

> **Intrinsic Existence Axiom**: consciousness exists as an intrinsic property, that is, my experience exists independently of external stimulus, as illustrated by hallucinations of external stimuli. Moreover, a conscious subject cannot doubt one's ongoing experiences, which are private and immediately known.
>
> **Composition Axiom**: consciousness is structured in that each experience is composed of many phenomenological distinctions. For example, an experience of a blue book may be broken down into an experience of blue, an experience of a rectangle, etc.

[2] See Tsuchiya et. al. (2017) and Haun et. al. (2016). However, the empirical case for IIT should not be overstated, and has been challenged e.g. in Barrett & Seth (2011), Peressini (2013), and Cerullo (2015: pp.4-5).

[3] "As recognized by Descartes, my own experience is the only thing whose existence is immediately and absolutely evident" (Tononi et. al. 2016: p451).

[4] The following list is based on Tononi and Koch (2016) and Tononi et. al. (2016).

**Information Axiom**: each conscious experience is the particular way it is by differentiating itself from what it is not. For example, an experience of a blue wall is what it is in part because it is not an experience of a green wall, etc. The more an experience rules out, the more information it contains.

**Integration Axiom** consciousness is unified in the sense that each experience is not reducible to non-interdependent subsets of phenomenological distinctions. For example, an experience of a blue book is not simply a colorless book-shape combined with disembodied blue. These phenomenological distinctions are integrated into a whole experience.

**Exclusion Axiom**: consciousness is definite in content and spatio-temporal grain: each experience has the set of phenomenological distinctions it has, neither less nor more, and it flows at the speed it flows, neither faster nor slower. For example, my experience is never a superposition of a visual field with one boundary and another visual field with a greater boundary. Nor does experience ever enter into a superposition of flowing at distinct speeds.

Many critics have taken issue with these axioms.[5] We can distinguish three types of problem that can be raised for them. Firstly, some axioms do not seem sufficiently precise. Secondly, some are sufficiently precise, but seem inaccurate. Thirdly, the axioms seem incomplete, or insufficient for fully capturing phenomenal consciousness. These are serious problems which I will simply flag now and address in section 6. For now, let us continue the exposition of IIT.

IIT now attempts to specify what properties a physical system must have in order to support consciousness. For this, a corresponding set of physical postulates is formulated. To describe these postulates, I will use a simple example that is based on the example in Tsuchiya (2017).[6] The example is described in *(a) Figure 1.* and will be referenced to help explain the postulates.

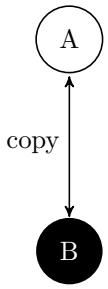
**Intrinsic Existence Postulate**: to support consciousness, a physical system must have *intrinsic* causal power. The system must have causal power *over itself*, independently of external factors. At minimum, this requires that there are some future states that the system can reach from some initial state with probability greater than chance, just in virtue of its internal structure. AB (figure 1) satisfies this postulate since its internal structure guarantees with certainty that it will reach state A=0; B=1 from state A=1; B=0.

**Composition Postulate**: to support consciousness, a physical system must be structured into parts that themselves have causal power within the system. Being made of neurons A and B, system AB satisfies this postulate.

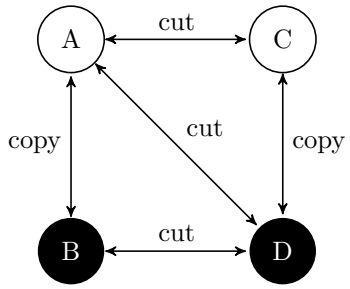
**Information Postulate**: to support consciousness, a physical system must specify a causal structure that differentiates its state at one time from its state at other times. AB satisfies this axiom since we can distinguish four possible present states (both on; both off; A=1,B=0; A=0,B=1). It is possible to measure the extent to which the system in a given state *constrains* its possible future states. For example, if AB is in state A=1, B=0, then AB contains *two units of information* about its immediate future state.

[5] For example Peressini (2013) and Cerullo (2015).

[6] For simplicity, I leave out some details that are inessential to the discussion. For example, I calculate the integrated information a system has about its future state. But IIT also requires that we calculate the integrated information the system has about its past state and take the minimum value. Furthermore, the mathematical formalism has undergone significant evolution (from IIT 1.0 to IIT 3.0). The present discussion is based on IIT 2.0, which is sufficient for our purposes




(a) ***Figure 1.*** **Two connected neurons, A and B.** They can be either off (black=0) or on (white=1). They function as copy gates such that (after some time lag) they turn connected neurons into their current state. The AB system has four possible initial states and is depicted as being in the specific initial state A=1, B=0. So given AB's internal causal powers, the next state will be A=0, B=1. We can say that when AB is in state A=1, B=0, AB contains intrinsic information I(AB) about its later state, since it cuts down *four* possible states into *one*. Logarithm base 2 gives units of information such that I(AB) = $\log_2(4)$ - $\log_2(1)$ = 2. So AB has two units of intrinsic information. This information is also integrated. For we would lose it by ignoring ("cutting") the connection between A and B: A by itself gives zero information about A's next state. Same with B. So the integrated information of AB, Φ(AB) = I(AB) - (I(A) + I(B)) = (2 - (0 + 0)). AB therefore has two units of consciousness.




(b) ***Figure 2.*** **An idealized split brain patient.** AB is the left hemisphere, CD is the right hemisphere. The connections between the two hemispheres have been cut. From *figure 1* we know that I(AB) = Φ(AB) = I(CD) = Φ(CD) = 2. The exclusion postulate then entails that the whole brain ABCD is only conscious if Φ(ABCD) > 2. But Φ(ABCD) = I(ABCD) - (I(AB) + I(CD)). That is, Φ(ABCD) = (4 - (2 + 2)) = 0. Hence, the split brain is not conscious, the hemispheres are. However, prior to the cuts, the exclusion postulate would entail only one conscious mind since in that case Φ(ABCD) > Φ(AB) and Φ(ABCD) > Φ(CD). This is IIT's explanation of split-brain cases.

**Integration Postulate**: to support consciousness, a physical system's causal structure must be unified, or irreducible to a simple sum of component causal structures. AB satisfies this axiom since it is the causal *connections* between A and B that constrain AB's possible future states. Ignoring those connections, thereby treating A and B as independent parts, prevents us from ruling out so many possible future states for AB. The extent to which a system's causal structure is irreducible in this way is measured by its amount of *integrated* information or Φ. Since A (and also B) carries zero information about its own future state, Φ(AB) = 2.

**Exclusion Postulate**: to support consciousness, the system's causal structure must be specified over a single set of elements, the set that yields the *maximum* amount of integrated information. For example, although our AB system has nonzero Φ, there are two situations in which it would exhibit zero consciousness. The first is if AB is a component of a bigger system (call it ABX) which itself has Φ > 2. If ABX is a closed system then although AB would contribute to ABX's consciousness, AB would not itself be conscious. Here we say that ABX has $\Phi^{max}$. The second situation is when AB contains a part that has Φ > 2. If AB is a closed system then although it has nonzero Φ, it would have no consciousness. Instead, its $\Phi^{max}$ part would be conscious.

A final aspect of the theory is the crucial distinction between the *quantity* of consciousness ($\Phi^{max}$), and the *quality* of a conscious experience. The latter is specified by the informational relationships generated by the conscious system (Balduzzi & Tononi (2009)). For any system in a given state, one can in principle draw up its informational relationships in a space called

"qualia-space". Every possible experiential quality corresponds to a particular shape composed of information relationships in this space. This is important for testing IIT. For in principle, one would like to deduce the exact phenomenal quality of a subject's experience from the informational shape determined by the subject's brain state.

Above I mentioned three problems for the IIT axioms: unclarity, inaccuracy, and incompleteness. There is a fourth problem for IIT, which concerns the connection between the axioms and the postulates. Granting the correctness of the axioms, it is not clear whether the postulates are even suggested by the axioms, let alone derivable from them. It will later be argued that all four of these problems are removed if we adopt illusionist integrated information theory.

To conclude the exposition of IIT let us illustrate its potential predictive and explanatory power using two examples. The first example concerns the fact that the unconscious cerebellum has far more neurons than the conscious cerebrum. This fact suggests neuron *number* is not relevant to consciousness whereas neuron *interaction* is. And indeed, cerebrum neurons are highly integrated whereas cerebellum neurons are not. Probe a region of the cerebellum and there is little disruption to the overall causal network of cerebellum neurons. But probe a region of the cerebrum and there is significant disruption to the overall causal network of cerebrum neurons. IIT can explain this by the fact that the cerebrum $\Phi^{max}$ is much greater than the cerebellum $\Phi^{max}$ (Tononi & Koch 2015: p10).

The second illustration of IIT's explanatory power concerns split-brain patients, and the fact that splitting the two brain hemispheres can generate two conscious minds that are continuous with the former unified mind. Any attempt to explain consciousness in terms of brain functions must explain this. According to IIT, prior to the connections between the two hemispheres being severed, both hemispheres have some $\Phi$, but only the whole brain has $\Phi^{max}$. Meanwhile after the connections are severed, the whole brain has less $\Phi$ than either of its hemispheres, and each hemisphere enjoys its own $\Phi^{max}$. This is illustrated in *(b) figure 2*. The exclusion postulate entails that if $\Phi(\text{ABCD}) < 2$, then ABCD has no consciousness at all. Rather, AB and CD each have their own conscious minds. This is how IIT explains split-brain patients (Tononi & Koch 2015: p10).[7]

IIT is a promising theory of consciousness. But it also faces significant problems. I have already briefly mentioned four problems regarding the axioms and postulates. But perhaps the deepest problem is the so-called hard problem of consciousness.

# 3 The hard problem of consciousness

The hard problem of consciousness is the problem of reductively explaining phenomenal consciousness in terms of physical processes.

According to Chalmers (2012: pp.307-8), a successful reductive explanation will give just enough detail to make it plausible that the explanandum is deducible from the explanans.[8] If we adopt this constraint for consciousness, then a successful reductive explanation of consciousness must make it plausible that consciousness is deducible from a physical description of the relevant features of conscious subjects. This is why Chalmers (1996) appeals to the logical possibility (or "conceivability") of zombies to argue that a reductive explanation of consciousness in physical terms is impossible. A zombie is a physical duplicate of oneself that does not have phenomenal

[7] The "two-streams" model of the split-brain is assumed in IIT's explanation. This model has been challenged e.g. in Bayne (2007, 2008). Note that Bayne (2007: p9) nonetheless describes it as "the received view" of the split-brain.

[8] McQueen (2015) argues that the success of reductive explanations in physics (specifically, reductive explanations of macrophysical properties in terms of microphysical descriptions) depend in part on the fact that they meet this constraint. See also Chalmers (2012: sec. 6.15).

consciousness. If the physical description of oneself is consistent with the absence of consciousness, then consciousness is not deducible from that physical description. But then no reductive explanation of consciousness in physical terms can be successful.

For the purposes of this section I will adopt this explanatory constraint as it will help to show why IIT exacerbates the hard problem. If IIT is to reductively explain consciousness, it must make plausible the claim that consciousness is deducible from its physical descriptions, thereby rendering zombies inconceivable. IIT's physical descriptions are given by its physical postulates, which describe the patterns of integrated information in conscious physical systems. IIT makes the hard problem look worse since we apparently have all relevant information available to *completely* describe the integrated information patterns of certain simple high-$\Phi$ systems, yet we cannot deduce consciousness from those descriptions. Non-conscious duplicates of such systems seem entirely conceivable. Let us consider some simple systems with nonzero $\Phi$.

The cerebrum has much greater $\Phi$ than the cerebellum. But the cerebellum will still have pockets of $\Phi^{max}$ spikes and will therefore have pockets of consciousness. This is allowed by the exclusion postulate. To see this, recall that the reason for why a split brain does not have one unified conscious mind is that each hemisphere has greater $\Phi$ than their union, so the consciousness of their union is excluded in favour of the consciousnesses of the hemispheres (*(b) Figure 2*). Arguably, the same thing is happening in the cerebellum, where pockets of neural mechanisms in the cerebellum will have greater $\Phi$ than their union, thereby excluding the consciousness of the cerebellum in favour of the consciousnesses of its components. We may expand these considerations to isolated molecules, which may have some consciousness in virtue of having nonzero $\Phi$ and being isolated from more complex systems. We can imagine a world that contains nothing but one such molecule. It is hard to see how we could deduce consciousness from the physical description of such a simple world. A "zombie-molecule" seems entirely conceivable.

In IIT, a system need not be as complex as a molecule to exhibit consciousness. Our simple isolated AB-system (*(a) Figure 1.*) was calculated to possess two units of $\Phi$ and so two units of consciousness. This is not nearly as much $\Phi$ as a human cerebrum (which, according to some estimates, has $\Phi$ equal to one billion). But AB is still conscious to a degree according to IIT, and it is rather difficult to see how we could deduce consciousness from AB's physical description alone. An "AB-zombie" is certainly conceivable.

Such examples were brought to their logical extreme when Scott Aaronson calculated that certain very simple systems can be constructed so as to have much more $\Phi$ than human brains. For example (Tononi, 2014), $N$ XOR gates arranged in a grid would have $\Phi^{max} = \sqrt{N}$. If we estimate the human cerebrum to have $\Phi^{max}$ equal to one billion, then stacking one billion squared XOR gates along a 2D grid would create a physical system with the same amount of consciousness as a human. It is very puzzling as to how such a system could gain so much consciousness just by adding more XOR gates to it. Moreover, a "grid zombie" is entirely conceivable: there is no logical inconsistency in there being nothing it is like to be such a grid.

Such examples mean that the standard response to the hard problem is unavailable to IIT. It is natural to respond that zombies may well be conceivable for us now, but that this is due to our current ignorance of the underlying physical details of the brain.[9] For example, Dennett (1996) argues that there is no more of a "hard" problem of consciousness for us now than there was a "hard" problem of life for seventeenth century vitalists. Likewise Block (2009: p1115), following Nagel (1974), argues that our situation is analogous to that of pre-Socratic philosophers who had no way of understanding how *heat* could be a kind of *motion*, because they lacked the appropriate concepts of motion (e.g. kinetic energy) that would allow an understanding of how

[9]See e.g. van Gulick 2000. This is described as *the* standard response by Worley (2003) and Brueckner (2001), and likely still is within the scientific community. Within more recent analytic philosophy of mind, it is possible that the so-called phenomenal concepts strategy (Stoljar 2005) has become more popular.

such different concepts could pick out the same phenomenon. The conscious grid blocks this response, since there is no relevant underlying details of the grid that we are ignorant of. We can fully understand why it generates such high $\Phi$, but this full understanding hardly suggests that grid-zombies are inconceivable. This is the sense in which IIT exacerbates the hard problem.

When considering the hard problem, Tononi and Koch (2015: p5) say this,

> "Indeed, as long as one starts from the brain and asks how it could possibly give rise to experience–in effect trying to 'distill' mind out of matter, the problem may be not only hard, but almost impossible to solve. But things may be less hard if one takes the opposite approach: start from consciousness itself, by identifying its essential properties, and ask what kinds of physical mechanisms could possibly account for them. This is the approach taken by [IIT]".

Here we should distinguish the context in which we *discover* theories from the context in which we *justify* theories [Schickore 2014: sec. 5]. I agree with Tononi and Koch that in the context of trying to discover a theory of consciousness, we should ask what kinds of physical mechanisms could account for consciousness. However, it is in the context of *justification* that the theory must make it plausible that the explanandum is deducible from the explanans. By IIT's own lights, we already have an effectively complete physical description of a grid of identical logic gates that has more $\Phi^{max}$ than a human. So we are now looking to *justify* IIT with respect to this example. The problem is that the absence of consciousness is entirely consistent with IIT's physical description of the grid.

To solve the hard problem we need to know exactly how consciousness is related to its physical correlates in the brain. This involves engaging philosophical theories of consciousness that specify the metaphysical relation binding consciousness to the physical world. Unfortunately, Tononi and Koch's remarks appear inconsistent on this matter. In some places they advocate Russellian pansychism, which states that consciousness is fundamental, and constitutes the intrinsic nature of physical matter (Tononi & Koch 2015: p11). In other places, they advocate physicalism by postulating an identity between consciousness and patterns of integrated information (Tononi & Koch 2015: p9). However, IIT is in tension with both philosophical theories. It is argued, quite persuasively, that IIT is logically inconsistent with Russellian panpsychism in (Mørch, forthcoming).[10] And here, we have argued that the conjunction of IIT and physicalism is at the very least incomplete.[11] So let's try something else.

# 4 From the hard problem to the illusion problem

There is a philosophy of mind which tries to side-step the hard problem entirely by denying that phenomenal consciousness exists. Recently, this position received a powerful defence by Frankish (2016a, 2016b), who refers to the idea as *illusionism*.[12]

Illusionism denies that experiences have phenomenal properties. Here 'experience' is defined functionally, as a mental state that is the direct output of sensory systems. The key claim of

[10] Russellian panpsychism requires that consciousness is a fundamental property of nature. Barrett & Seth (2011), Peressini (2013: sec. 2.1), and Barrett (2014) all offer arguments against the claim that integrated information could be, or could be exactly correlated with, a fundamental property. I discuss these in section 6.

[11] Mindt (2017) also argues that the conjunction of physicalism and IIT yields an incomplete theory. He argues that IIT only specifies structure and dynamics and that specifying only structure and dynamics is insufficient to explain consciousness.

[12] Dennett (1988, 1991) pioneered a similar position, which is often referred to as *eliminativism*. According to Dennett (2016: 65), among philosophical theories of consciousness, "illusionism as articulated by Frankish should be considered the front runner".

illusionism is that experiences have *quasi-phenomenal properties*. A quasi-phenomenal property is a non-phenomenal, physical property of experience that introspection typically misrepresents as phenomenal. In other words, introspection represents quasi-phenomenal properties as if there is something it is like to have them, when in fact there is not.

Frankish does not offer much in the way of an account of introspection, except to say that introspection issues in dispositions to make phenomenal judgments (judgments about the phenomenal character of particular experiences and about phenomenal consciousness in general), either directly or by generating intermediate representations of sensory states which ground our phenomenal judgments (2016a: p14). According to Beaton (2009: p7), introspection is identical to the ability to gain knowledge in a fundamentally first-person way; even if the relevant knowledge is not gained entirely through introspection (in this sense) it must be gained in a way which essentially involves introspection. Beaton (2009: p15) adds that on essentially any theory of introspection, we can introspect "propositional attitude-style states", including "seeing $x$" and "experiencing $x$", where $x$ is some possible public state of affairs. For example, introspection is what allows the right kind of subject to know *that* she is a seeing a red ball when she is seeing a red ball.[13] Introspection may generate the belief that *there is something it is like* to see the red ball, such that this something is physically inexplicable. According to illusionism, this would be an introspective illusion.

If illusionism is true, there is no hard problem of consciousness. Instead, there is the *illusion problem*. The illusion problem is the problem of explaining why experiences seem to have phenomenal properties (when they in fact don't). Frankish (2016a: pp.27-29) offers several arguments in favour of illusionism. But the main argument seems to be that the illusion problem is in principle solvable whereas the hard problem is not. For illusionism is not forced to explain how physical processes give rise to phenomenal consciousness. Instead, it need only explain how physical systems like us misrepresent themselves as having phenomenal properties. Arguably, this looks more tractable. Indeed in section 7 I will argue that it is an instance of an "easy" problem of consciousness since it only involves the mechanistic explanation of cognitive functions.

It is also arguable that when we focus on illusionism, we can better unify our theories of consciousness with the rest of science. For example Humphrey (2011) connects illusionism with evolutionary theory. We may be able to better integrate consciousness science with evolutionary psychology if we start thinking of evolutionary explanations for why we have quasi-phenomenal properties. Perhaps it was highly adaptive to represent your brain states as having apparently non-physical properties. Perhaps this helped to give our ancestor's lives meanings, thereby motivating them into more advanced activities.

To illustrate how illusionism removes the hard problem, it is instructive to consider what illusionism says about the conceivability of zombies. A zombie is defined as an exact physical duplicate of oneself that is not phenomenally conscious. Illusionism trivializes the conceivability of zombies: we are zombies and we can conceive of ourselves therefore, zombies are conceivable.

Given the way that zombies are often characterized, it might sound outrageous to say that we are zombies. For sometimes zombies are said to have no subjectivity, and no internal inner life such that there is nothing it is like to be them and they are "all dark" inside (Chalmers, 1996: pp.95-6). But given illusionism, this is a misleading description of zombies. For a physical duplicate of oneself will have quasi-phenomenal properties that are introspectively misrepresented, and these misrepresentations will cause the duplicate to assert that it is phenomenally conscious. So there is at least a sense in which zombies are dramatically distinct from inanimate objects. To help capture this difference, Frankish (2016a: 23) suggests that we disambiguate a second notion of "what it is like", where we can say that there is something it is like to be in a state if that

[13]The difficulties in making introspection more precise are notorious (see e.g. Prinz (2004) and Schwitzgebel (2016)). But this is a reasonable starting point.

state is introspectively (mis)represented as having phenomenal properties. Given the importance of "what it is like" talk in human story-telling, among other things, this disambiguation seems essential.

Given illusionism, there is no need to physically explain phenomenal properties, and so there is no hard problem. However, there is a need to physically explain quasi-phenomenal properties. This is the illusion problem and it is undeniably formidable. Frankish (2016a: pp.29-37) notes that a number of challenges must be overcome if we are to solve the illusion problem. Here I wish to focus on one particularly difficult aspect of the illusion problem, which Frankish (2016a: p35) attributes to Levine (2001: pp.146–7), and calls the problem of *representing phenomenality*:

> "If there are no phenomenal properties, how do we represent them? How do we acquire phenomenal concepts and how do these concepts capture the richness of phenomenality? These are central questions for illusionists, and answering them would go a long way towards solving the illusion problem."

To elaborate, even if a concept fails to pick out anything in the world, it is bound to have some content that can be found in the world. If it didn't, it would be difficult to see how we could have acquired the concept in the first place. As a simple example, our concept *witch* has been found to not refer to anything. But the concept is made up of parts such that some of those parts do correspond to things we find in reality, concepts such as *woman*, *broomstick*, *pointy hat*, etc. Acquisition of the concept can then be explained in terms of confused thinking that wrongly pulled these (and other) concepts together. The same occurs in more serious scientific examples. The concept *aether* refers to material that fills the region of the universe above the terrestrial sphere and which explains the traveling of light and gravity. Nothing in reality corresponds to this concept. But plenty corresponds to the concepts used to define *aether*, such as light and gravity. Acquisition of the concept can then be explained in terms of confused thinking about the causes of the behaviour of light and gravity. The problem with phenomenal concepts is that *the entire category of being* to which they refer seems not to exist. How, then, could we have possibly acquired them?

Towards a solution, Frankish offers some "preliminary" remarks, which are intended to "indicate some lines open to the illusionist." Frankish devotes most of this discussion to the idea that phenomenal concepts are *hybrid* concepts. Some concepts are theoretical concepts, they apply to something defined in terms of potentially false theories (as with *aether*). Other concepts are purely recognitional, they apply to something on the basis of recognition. Simple colour concepts might be examples of recognitional concepts: I apply *red* to a surface not because my theory tells me that the surface is red, but because the surface disposes me to call it red. A hybrid concept is both theoretical and recognitional. Frankish's idea is that some complex physical property is recognised by introspection as being "that type of phenomenal property", where "phenomenal property" is the theoretical component that treats the complex property as intrinsic, immediately known, non-physical, etc.

I think the hybrid concept idea is a good start. But much more needs to be said about what the relevant complex physical properties are, how they could be recognized by introspection, and what it means for phenomenal concepts to represent those properties as phenomenal. If we could at least specify the complex physical properties that phenomenal concepts apply to, then we would progress for at least two reasons. Firstly, this would help specify the veridical aspects of phenomenal concepts. From there, we might be able to explain the acquisition of phenomenal concepts in terms of some sort of confused thinking about the veridical aspects of their content. Secondly, this would help give us the neurophysiological states that introspection acts upon, thereby enabling further progress in the scientific study of introspection.

In what follows I will define a theory - illusionist-IIT - which specifies a natural way of combining IIT and illusionism: the veridical content of hybrid phenomenal concepts is integrated information. I then show how illusionist-IIT can help resolve outstanding problems for IIT (section 6). I will then try to make progress on the illusion problem.

# 5 Illusionist Integrated Information Theory

On the face of it, illusionism and IIT are incompatible. This isn't just because IIT is formulated as a theory of something whose very existence is denied by illusionism. It is also because the general approaches are so different: whereas illusionism proceeds by treating introspection as defective, IIT proceeds by treating introspection as sacrosanct and providing the primary evidence for the theory. However, these approaches are not mandatory.

In the case of IIT, it is rather implausible that the primary evidence for IIT is purely phenomenological, whereas the third-person evidence is secondary. More plausibly, third-person evidence (e.g. cerebrum/cerebellum comparison) that support the idea of a complexity measure of consciousness comes first. Then, phenomenological considerations (the axioms) simply help to refine the complexity measure, and to justify why it should measure something like integrated information in particular. Finally, further third-person tests, e.g. the earlier mentioned experiments that predict subject's experiences based on their integrated information patterns, complete the empirical justification of IIT. Looking at the justification of IIT in this way enables us to lighten the heavy epistemic burden that Tononi and others have placed on introspection. And in the case of illusionism, it is sufficient if only certain components of phenomenal concepts are non-veridical. This allows that some components (those suggestive of integrated information) are partially veridical.

Let us then reformulate the axioms. They can no longer be axioms about phenomenal consciousness (which does not exist, according to illusionism). Instead, they should be axioms about quasi-phenomenal properties, and how they are introspectively represented. First we need a new axiom, the illusion axiom, that asserts the existence of quasi-phenomenal properties. Here are the revised axioms:

> **Illusion Axiom**: experiences have quasi-phenomenal properties, where a quasi-phenomenal property is a non-phenomenal property of experience that introspection typically misrepresents as phenomenal.
>
> **Intrinsic Existence Axiom**: quasi-phenomenal properties are introspectively represented as being intrinsic properties that are private and immediately known.
>
> **Composition Axiom**: quasi-phenomenal properties are introspectively represented as being composed of many phenomenological distinctions.
>
> **Information Axiom**: quasi-phenomenal properties are introspectively represented in terms of informative phenomenological differences. That is, each experience appears to be the particular way it is by differentiating itself from what it is not.
>
> **Integration Axiom**: quasi-phenomenal properties are introspectively represented as being unified in the sense that each experience is irreducible to non-interdependent subsets of phenomenological distinctions.
>
> **Exclusion Axiom**: quasi-phenomenal properties are introspectively represented as being definite in content and spatio-temporal grain.

Given these axioms we can now try to specify what properties a physical system must have in order to support quasi-phenomenal properties. Let us begin with the postulate that explains the illusion axiom:

> **Illusion Postulate**: to support quasi-phenomenal properties, a system must have the type of physical property that is systematically introspectively misrepresented as being phenomenal.

If we take the strict illusionist stance that phenomenal concepts are entirely non-veridical, then it would seem that this postulate is all we can get from the axioms. For the postulates should be about the physical structure of quasi-phenomenal properties (the represented system). But if the representations specified in the axioms all misrepresent the structure of quasi-phenomenal properties, then the real postulates need look nothing like the axioms.

However, I have argued that it is difficult to understand phenomenal concepts if they are entirely non-veridical. In light of this, the best way to proceed is to assume as much as possible that the representations specified in the axioms accurately represent quasi-phenomenal properties. If the resulting postulates specify empirically plausible physical correlates of consciousness, then we were right to treat the axioms veridically. However, if they do not, then we can blame that on the axioms specifying nonveridical representations, and revise the postulates accordingly. This seems like a more cautious and more scientific approach than insisting on axioms that specify "self-evident truths about consciousness – the only truths that, with Descartes, cannot be doubted and do not need proof" (Oizumi et. al. 2014: 2). It is also a promising method for isolating the accurate content of phenomenal concepts (if any), to help solve the problem of representing phenomenality.

As discussed in section 2, there are reasons to think that IIT specifies empirically plausible physical correlates of consciousness. In that case, let's take all the axioms at face value, and adopt all five of the IIT postulates. Of course, the new postulates cannot specify properties that a system must have for that system to support phenomenal consciousness. Instead, they must be formulated as specifying properties that a system must have for that system to support quasi-phenomenal properties. The postulates then become:

> **Intrinsic Existence Postulate**: to support quasi-phenomenal properties, the system must have intrinsic causal power. This means having causal power over itself, independently of external factors.
>
> **Composition Postulate**: to support quasi-phenomenal properties, the system must be structured into parts that themselves have causal power within the system.
>
> **Information Postulate**: to support quasi-phenomenal properties, the system must specify a causal structure that differentiates its state at one time from its state at other times. That is, it must contain information about itself.
>
> **Integration Postulate**: to support quasi-phenomenal properties, the system must be unified, or irreducible to a simple sum of component causal structures. That is, the self-information it contains must be integrated.
>
> **Exclusion Postulate**: to support quasi-phenomenal properties, the system's causal structure must be specified over a single set of elements, the set that yields the maxmimum amount of integrated information.

Let us now put illusionist-IIT to work.

# 6 Problem solving with Illusionist-IIT

I have so far identified five problems for IIT: (i) axiom unclarity; (ii) axiom inaccuracy; (iii) axiom insufficiency; (iv) axiom-postulate disparity; and (v) conscious logic gates. In this section I will elaborate on each and argue that illusionist-IIT makes progress on solving them. I will also consider how illusionist-IIT removes some additional technical problems that concern the formal definition of integrated information (vi). Finally, I consider (vii) the hard problem of consciousness.

(i) *Axiom unclarity*: the content of some IIT axioms are unclear. For example, the intrinsic existence axiom typically contains a metaphysical claim and an epistemic claim (Tononi and Koch, 2015: p5; Tononi et. al. 2016: p451). The metaphysical claim is that phenomenal properties are metaphysically intrinsic. The metaphysical claim is unclear as there are recent challenges to the very concept of metaphysical intrinsicness (McQueen and van Woudenberg, 2016). Illusionist-IIT removes this unclarity since quasi-phenomenal properties are only defined as being *represented* as intrinsic. This could be a misrepresentation that involves defective concepts. If it is, then while IIT's intrinsic existence axiom would be false, Illusionist-IIT's still holds and helps to specify the representational content of phenomenal concepts. This may force us to revise the corresponding postulate. But since the postulate translates metaphysical intrinsicness into internal causal power (a less controversial concept), this is not necessary.[14]

(ii) *Axiom inaccuracy*: IIT's intrinsic existence axiom also specifies an epistemic claim: phenomenal properties are private and known with immediate certainty. It is not clear that this is accurate. It is not a priori since it is conceivable that the best explanation of introspection involves denying its reliability (See e.g. Dennett (1988, 1991) and Beaton 2009)). Moreover, introspection is routinely found to create illusions. Subjects systematically judge that their visual field is only blurry at the outer fringes, until they are asked to focus their vision on a point in front of them, slowly bring a playing card towards that focal point, and say what card it is before it reaches the focal point (Westerhoff, 2010: p20). The IIT axioms are also just introspective judgments that are corrigible in light of the unreliability of introspection. Illusionist-IIT resolves this by removing any mention of epistemic certainty from the axioms.

(iii) *Axiom insufficiency*: IIT specifies five putative facts about phenomenal consciousness, but gives no proof that these are the only relevant facts to be considered. The axioms may therefore specify necessary but not sufficient conditions for consciousness. This seems especially plausible when we reflect on what IIT implies for certain simple systems such as the AB-system (*(a) Figure 1*). It is reasonable to suppose that any theory that renders AB conscious is yet to specify sufficient conditions for consciousness. Illusionist-IIT removes this problem by moving the focus from phenomenal properties to quasi-phenomenal properties. The Illusion axiom states that quasi-phenomenal properties are typically misrepresented by introspection. In that case, we no longer have counterintuitive predictions: although the AB-system has the kind of property that is typically misrepresented by introspection, the AB-system is not actually capable of introspection. Hence, there is no sense in which there is something it is like to be the AB-system. On the illusionist's deflationary notion of "what it is like", a system must be capable of introspection if there is going to be something it is like to be that system.

(iv) *Axiom-postulate disparity*: some IIT postulates appear to only vaguely resemble their corresponding axioms. For example compare the information postulate with the information axiom. The axiom states that experiences are partly characterized in terms of their differences:

[14]To add to the problem of unclarity, some presentations of IIT's intrinsic existence axiom skip the metaphysical component entirely, and only include the epistemic component e.g. Oizumi, Albantakis, & Tononi (2014: p2). This also exacerbates the axiom-postulate disparity problem since the epistemic component seems unrelated to IIT's intrinsic existence postulate.

pain is what it is partly in virtue of not being pleasure, a red experience is what it is partly in virtue of not being green or blue. There is therefore a sense in which an experience can contain information by ruling out possibilities. But these possibilities are *synchronic*: the experience yields information about the present by ruling out other states as *not* being present. But the possibilities specified in the information postulate are *diachronic*: the physical state yields information about the future (or past) by ruling out possible future (or past) states. So, it is unclear why the information postulate supports the information axiom since they concern quite distinct modal constraints. Illusionist-IIT avoids this problem by allowing that the representations specified in its axioms are only approximately correct. For example, introspection is right to a represent a modal constraint, but gets the exact kind of constraint wrong.

(v) *The logic gate problem*: IIT entails that a 2D grid of around one billion squared identical XOR gates would have as much $\Phi$ and therefore as much consciousness as a human. Illusionist-IIT entails that nothing is phenomenally conscious and so trivially entails that the grid is not phenomenally conscious. More importantly, the quasi-phenomenal properties of the grids are not introspectively represented. The grids were designed only to maximize $\Phi$. They have not, in addition, been designed to support a complex introspective mechanism. Thus, while there is something it is like to be us (in the previously defined sense that we undergo introspective illusions) there is nothing it is like to be the grid (in *any* relevant sense).

(vi) *Technical problems*: As mentioned above (note 10), Tononi treats consciousness as a fundamental property. Integrated information therefore is, or at least perfectly correlates with, a fundamental property. However, several authors have argued that integrated information is not up to this task. Thus, Barrett and Seth (2011) offer what they take to be improved measures of integrated information. However, their measures are not invariant under changes in coordinates (p14) whereas fundamental properties are usually required to be invariant under such changes. Peressini (2013: sec. 2.1) argues that for some systems there may be no unique decomposition of a system into subsystems, meaning there is no unique minimum information partition to define a unique value for $\phi$. In addition, there may be no unique way of stating the causal relationships between its subsystems, meaning there is no unique way to define a system's information. Finally, Barrett (2014) argues that existing formulations of IIT are not applicable to standard models of fundamental physical entities since standard models describe such entities in terms of continuous fields. In rejecting the idea that consciousness is fundamental, illusionist-IIT evades these objections. In particular, integrated information may be conceived as a non-fundamental, frame variant, sometimes non-uniquely defined property of discrete systems. What's important is that integrated information has certain higher-level causal powers that make it prone to being misrepresented by introspection. Thus, illusionist-IIT predicts that introspection is causally sensitive to integrated information among neurons and groups of neural mechanisms. Introspection need not be sensitive to integrated information that obtains at the level of elementary particles or continuous fields.

(vii) *The hard problem*: the hard problem is the problem of physically explaining the existence of phenomenal consciousness. This problem is removed if phenomenal consciousness does not exist. By taking the illusionist stance, we replace the hard problem with the illusion problem: the problem of explaining how introspection systematically misrepresents experiences as phenomenal. The illusion problem is multi-faceted and remains unresolved. In the next section I argue that illusionist-IIT makes progress on some aspects of the problem.

# 7 Towards a solution to the illusion problem

If there is an empirical case for integrated information being the physical correlate of phenomenal consciousness, then there is a corresponding puzzle for illusionism: if phenomenal consciousness does not exist, why does it appear to have a stable physical correlate in the brain? Illusionist-IIT answers this question by developing an account of introspection according to which introspection monitors large-$\Phi$ states and creates representations of them - phenomenal concepts - whose non-veridical content causes us to believe in phenomenal consciousness.

One might object that any materialist theory that postulates some physical correlate of consciousness could be rephrased in terms of illusionism. This would involve stating that the proposed correlate is simply introspectively misrepresented as being phenomenal. Illusionist-IIT is distinctive since it helps to solve several problems that plague IIT, as argued in the previous section. But can illusionist-IIT also solve problems that plague illusionism? In particular, can it make progress on the illusion problem?

Solving the illusion problem will require more than just philosophical analysis, it will require significant empirical research too. As Marinsek and Gazzaniga (2016) emphasize in their response to Frankish, "a major limitation of illusionism is that it does not offer any mechanisms for how the illusion of phenomenal feelings works". One can speculate about various possible mechanisms, but ultimately we would like such proposals to be experimentally testable. Towards a solution, Chalmers (2018) argues that we can create computational models that build in versions of proposed mechanisms, and we can see whether these models reproduce something along the lines of human phenomenal reports, and he points to some preliminary research of this kind in Muehlhauser (2017). Chalmers goes on to argue that by using *principled* underlying mechanisms, we can attempt to build increasingly sophisticated systems that exhibit human-like phenomenal reports with increasing scope and accuracy. If it is possible to build a reasonably accurate system of this sort, the mechanisms it uses may provide a solution. The key word here is "principled".

In the remainder of the paper I suggest that illusionist-IIT is a framework for directing empirical research towards principled mechanisms. I return to the aspect of the illusion problem discussed in section 4, the problem representing phenomenality. I break this problem down into three more specific problems and show how illusionist-IIT creates research avenues for solving them. Three central questions that illusionists must answer are: (i) what *exactly* are the conceptual components of phenomenal concepts? (ii) why do we use such concepts rather than other representations? (iii) How can we explain the strength of the introspective illusion?

(i) *What are the conceptual components of phenomenal concepts?* Sections 4 and 5 offered a partial answer to this question. Recall that when a concept fails to pick out anything in the world, it will typically have at least some content that can be found in the world. This will then help to explain how we could have acquired the concept in the first place, and why it fails to refer. For example, the concept *phlogiston* refers to a fire-like element within combustible bodies that is released during combustion. Nothing in reality corresponds to this concept. But plenty corresponds to the component concepts used to define *phlogiston*, like *fire* and *combustion*. The acquisition of *phlogiston* can then be explained in terms of confused thinking about the causes of combustion.

Illusionist-IIT paves the way for this type of explanation for how we acquired phenomenal concepts. As discussed in section 5, IIT's consciousness axioms can be viewed as specifying components of phenomenal concepts (e.g. *being integrated*). IIT's postulates then state what would be required of a physical system if that content were veridical. If those postulates are empirically found to be satisfied in the brain, then we may conclude that we have discovered a veridical component of phenomenal concepts. Any axioms of consciousness that we specify,

that cannot be supported by the kinds of physical systems that correlate with consciousness, can then be treated as spelling out nonveridical components of phenomenal concepts. For example, Descartes thought it was essential to consciousness that it be non-extended, or non-spatial. We might therefore formulate an axiom stating that quasi-phenomenal properties are introspectively represented as being non-spatial. This would then be a prime candidate for a non-veridical component of phenomenal concepts.

Discovering axioms that correctly describe the components of phenomenal concepts is difficult. What happens when we disagree on what the axioms should be? It may well be that the structure of our phenomenal concepts differ (between individuals and cultures) with respect to their nonveridical components. This is an empirical question that can be determined by experimental psychology and experimental philosophy, e.g. by finding how widely shared are the various kinds of Chalmers' (2018) *problem intuitions*. These are intuitions underlying the belief that consciousness cannot be physically explained. Philosophical analysis will be essential to determine whether such intuitions reflect components of phenomenal concepts or just central beliefs about what those concepts refer to. Whatever the ultimate components of phenomenal concepts turn out to be, the point I wish to make here is that Illusionist-IIT provides empirically defensible proposals for the *veridical* components of phenomenal concepts.

(ii) *Why do we use such concepts rather than other representations?* One way of answering this question is to consider why it might have been inevitable that we evolve non-veridical representations of experiences. First, it would have been useful to evolve internal representations of experiences, for example, to communicate how things appear to one, how one is feeling, etc. Secondly, if materialism is true, then those representations represent extraordinarily complex physical states of the brain. In that case, it is likely that we would have evolved *novel* introspective representations that distort what is being represented. This line is developed by Chalmers (2018: sec. 2). But the question remains as to why we would have evolved one set of novel representations over another. Here illusionist-IIT requires that the form of the novel representations is at least constrained by the form of what they represent. In particular, reflection on those representations enable one to formulate axioms, whose corresponding postulates are in fact part of what is represented. Such a view may then be supplemented by the more speculative evolutionary considerations of Humphrey (2011), who argues that representing our internal states as having non-physical properties would have been fitness enhancing, since believing our minds are nonphysical makes our lives seem more meaningful and worth living.

*(iii) How can we explain the strength of the introspective illusion?* I will conclude by considering a potentially testable prediction of illusionist-IIT that relates the strength of the introspective illusion to $\phi^{max}$.

IIT predicts that the amount of phenomenal consciousness in an experience is proportional to its $\phi^{max}$. Hence, one way of experimentally testing IIT is by correlating the $\phi^{max}$ of brains of subjects with their judgments about the amount of phenomenal consciousness they experience. However, this is difficult in practice, not just because measuring the $\phi^{max}$ of a human brain is difficult, but also because the idea of *amount* of phenomenal consciousness is unclear.

Illusionist-IIT denies phenomenal consciousness and so must make different predictions. Here is a prediction of illusionist-IIT: *the strength of a subject's introspective illusion is proportional to the $\phi^{max}$ of the introspected state.* Hence, one way of experimentally testing illusionist-IIT is by correlating the $\phi^{max}$ of brains of subjects with the strength of their *problem intuitions*. This also raises practical concerns, for example, in determining whether a subject's expressed intuitions result from theory-neutral phenomenological reflection or from background theories of consciousness. Either way, it seems that illusionist-IIT is at least in principle experimentally distinguishable from IIT.

This prediction requires the existence of an introspective mechanism that is sensitive to

$\phi^{max}$. At this stage one can only speculate as to how this mechanism could work, for we are currently far from having a noncontroversial mechanistic account of how introspection could be sensitive to anything. But one thought is that for introspection to be sensitive to $\phi^{max}$, then it must be sensitive to some activity that comes along with, say, synchronization between remote neural networks, such as rapidly propagated activity through electrically coupled neurons. In principle this could be something that we could find in the brain: something that is sensitive to such phenomena and that in turn causes phenomenal judgments that are more likely to express problem intuitions.

# 8 Conclusion

IIT is a promising scientific theory of consciousness. However, it faces a number of problems concerning its axioms, postulates, and predictions, as well as its inability to address the hard problem. Illusionism is a promising philosophical theory of consciousness. However, it leaves behind the problem of explaining how introspection misrepresents states as being phenomenal when they are in fact not. This is a significant problem since it is hard to see how we could have acquired concepts that refer to a completely non-existent category of being. Here it has been argued that progress can be made on both fronts by revising IIT so that it can be appropriately combined with illusionism, resulting in what I have called *illusionist-IIT*.

Illusionist-IIT enjoys a number of advantages over IIT. It no longer faces the hard problem since illusionism replaces the hard problem with the illusion problem. The simplistic systems that IIT makes absurd predictions are not capable of introspection so do not have internal lives in the sense that humans do. And finally, the theory's physical postulates are no longer restricted by a set of immutable "consciousness axioms". Instead, the axioms are converted into statements about how introspected states are (potentially non-veridically) represented.

Illusionist-IIT also helps illusionism make progress on the outstanding illusion problem: the problem of explaining how introspection misrepresents states as being phenomenal. Illusionist-IIT provides us with the physical states that introspection represents and gives us a way of understanding how introspective representations are at least partially veridical. It is then possible to explain phenomenal concepts in terms of defective thinking about those physical states. This then suggests new cognitive mechanisms that could be potentially explained neurophysiologically, rendering the illusion problem an "easy" problem of consciousness.

Whether either IIT or illusionism are on the right track is an open question. But I hope to have made the case that there is more progress to be made in their combined illusionist-IIT form.[15]

[15]I would like to thank Hedda Hassel Mørch, Keith Hankins, Michael Pace, Michael Robinson, David Chalmers, and attendees of the Fall 2017 Metaphysics of Consciousness Seminar at Chapman University, for useful feedback and discussion.

# References


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

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

Bayne, T. (2007). Conscious States and Conscious Creatures: Explanation in the Scientific Study of Consciousness. *Philosophical Perspectives*, 21(1), 1–22.

Bayne, T. (2008). The Unity of Consciousness and the Split-Brain Syndrome. *Journal of Philosophy*, 105(6), 277-300.

Block, N. (2009). Comparing the Major Theories of Consciousness, in Gazzaniga, M. (ed.) *The Cognitive Neurosciences IV*. Cambridge, MA: MIT Press.

Brueckner, A. (2001). Chalmers's conceivability argument for dualism. *Analysis* 61: 187–93.

Cerullo, M.A. (2015). The Problem with Phi: A Critique of Integrated Information Theory. *PloS Computational Biology* 11(9): e1004286.

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

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

Chalmers, D.J. (2012). *Constructing the World*. Oxford: Oxford University Press.

Chalmers, D.J. (2018). *The Meta-Problem of Consciousness*. https://philarchive.org/archive/CHATMO-32. Accessed: 5/17/2018.

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

Dennett, D.C. (1996). Facing backwards on the problem of consciousness. *Journal of Consciousness Studies*, 3:4-6.

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

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

Frankish, K. (2016b). Not Disillusioned: Reply to Commentators. *Journal of Consciousness Studies*, 23(11-12): 256-289.

Haun, A., Kawasaki, H., Kovach, C., Oya, H., Howard, M. A., Adolphs, R., & Tsuchiya, N. (2016). Contents of Consciousness Investigated as Integrated Information in Direct Human Brain Recordings. *bioRxiv*. doi:10.1101/039032.

Humphrey, N. (2011). *Soul Dust: The Magic of Consciousness*. Princeton: Princeton University Press.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*, Oxford: Oxford University Press.

Marinsek, N.L. & Gazzaniga, M.S., (2016). A Split-Brain Perspective on Illusionism. *Journal of Consciousness Studies*, (23)11-12: 149-159.

McQueen, K.J. (2015). Mass Additivity and A Priori Entailment. *Synthese*, 192(5):1373-1392.

McQueen, K.J. & van Woudenberg, R. (2016). Tests for Intrinsicness Tested. *Philosophical Studies*, 173: 2935-2950.

Mindt, G. (2017). The Problem with the 'Information' in Integrated Information Theory. *Journal of Consciousness Studies*, 24(7-8): 130-54.

Mørch, H.H. (forthcoming). Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism? *Erkenntnis*.

Muehlhauser, L. (2017). *A Software Agent Illustrating Some Features of an Illusionist Account of Consciousness*. OpenPhilanthropy. [https://www.openphilanthropy.org/software-agentillustrating-some-features-illusionist-account-consciousness]

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.

Prinz, J.J. (2004). The Fractionation of Introspection. *Journal of Consciousness Studies*, 11(7–8): 40–57.

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

Schickore, Jutta. (2014). Scientific Discovery. *The Stanford Encyclopedia of Philosophy*, (Spring 2014 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2014/entries/scientific-discovery/.

Schwitzgebel, Eric. (2016). Introspection. *The Stanford Encyclopedia of Philosophy*, (Winter 2016 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/win2016/entries/introspection/.

Stoljar, D. (2005). Physicalism and Phenomenal Concepts. Mind and Language, 20: 469-94.

Tegmark M (2016). Improved Measures of Integrated Information. arXiv:1601.02626.

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. (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., & 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.

van Gulick, R. (1999). Conceiving beyond our means: The limits of thought experiments. In (S. Hameroff, A. Kaszniak, & D. Chalmers, eds) *Toward a Science of Consciousness III.* MIT Press.

Westerhoff J. (2010). *Twelve Examples of Illusion.* Oxford: Oxford University Press.

Worley, S. (2003). Conceivability, possibility and physicalism. *Analysis* 63:15-23.