This is a preprint version of a paper published in the Review of Philosophy and Psychology (https://link.springer.com/article/10.1007/s13164-022-00653-x). Please cite the published version

Can the Integrated Information Theory Explain Consciousness from Consciousness Itself?

Niccolò Negro

Department of Philosophy, Monash University, Australia

Abstract:

In consciousness science, theories often differ not only in the account of consciousness they arrive at, but also with respect to how they understand their starting point. Some approaches begin with experimentally gathered data, whereas others begin with phenomenologically gathered data. In this paper, I analyse how the most influential phenomenology-first approach, namely the Integrated Information Theory (IIT) of consciousness, fits its phenomenologically gathered data with explanatory hypotheses. First, I show that experimentally driven approaches hit an explanatory roadblock, since we cannot tell, at the present stage, which model of consciousness is best. Then, I show that IIT's phenomenology-first approach implies a self-evidencing explanation according to which consciousness can be explained by starting from consciousness itself. I claim that IIT can take advantage of the virtuous circularity of this reasoning, but it also faces a data-fitting issue that is somehow

similar to that faced by experiment-driven approaches: we are not given enough information to decide whether the explanatory hypotheses IIT employs to explain its phenomenological data are in fact best. I call this problem "the self-evidencing problem" for IIT, and after introducing it, I propose a possible way for IIT to solve it.

Keywords: Integrated Information Theory; Consciousness; Self-evidencing explanations; Phenomenology-first.

Introduction

In consciousness science, theories often differ not only in the account of consciousness they arrive at, but also with respect to how they understand their starting point. The main difference consists in whether a research programme takes first-person data (i.e. data gathered from one's own phenomenology) or third-person data (i.e. data gathered from objective methods such as behavioural and neurophysiological measures) as fundamental observations for theory-building (Chalmers, 2004).

In the philosophical and neuroscientific literature, attention has been mostly focused on how to best fit behavioural and neuroscientific data, and their correlations, with models of consciousness (Block, 2007; Chalmers, 2000, 2004; Fink, 2016; Hohwy & Frith, 2004). In this context, the problem is that different theories of consciousness provide different, incompatible, reasons as to why one model should be the best fit of experimentally gathered data.

In this paper, I focus instead on the "phenomenology-first" approach of the integrated information theory of consciousness (IIT) (Oizumi, Albantakis, & Tononi, 2014; Tononi, Boly, Massimini, & Koch, 2016) to determine whether this alternative approach is able to move the debate forward. I argue that IIT is affected by a similar, but importantly peculiar,

data-fitting issue: it is not clear how phenomenologically gathered data provide evidence for the particular explanatory model IIT provides.

This is a pressing issue for IIT, as its central claim is that there is an "explanatory identity" (Haun & Tononi, 2019, p. 5) between consciousness and integrated information. This identity depends on the fact that the integrated information measure derives from phenomenological evidence. So, integrated information is claimed to be a measure *of consciousness* because it is extracted *from* consciousness itself: the goal of IIT's phenomenology-first approach is then to explain consciousness from consciousness itself. But if there is no clear path that goes from phenomenology to integrated information, then the explanatory power of the theory *as a theory of consciousness* is undermined. The main claim of this paper is that even if we accept IIT's phenomenology-first approach, we still do not seem to have enough information to evaluate whether there is such a clear path; that is, we cannot determine whether the explanatory model proposed by IIT is in fact the best fit for the phenomenological evidence from which the theory starts: it is not clear that IIT succeeds in explaining consciousness from consciousness itself. I will call this "the self-evidencing problem" for IIT.

In the first section, I introduce the distinctive feature of IIT's approach by contrasting it with standard, experiment-driven, approaches in the neuroscience of consciousness. In the second section, I focus on how IIT builds a theory of consciousness from phenomenology, and specifically on the logic employed to infer explanatory hypotheses from phenomenological data. In the third section, I introduce self-evidencing explanations, and suggest that IIT could take advantage of this form of reasoning for "bootstrapping" the explanation of consciousness from consciousness itself. However, in section four, I show that for this strategy to work, IIT has to provide information not yet made explicit: IIT has a self-evidencing problem. After specifying the peculiar nature of the self-evidencing problem in section five, I propose a possible way forward for solving the self-evidencing problem, in section six. In the

conclusion, I point out that solving the self-evidencing problem would be crucial to IIT, as it is a necessary step in demonstrating the explanatory superiority of IIT's phenomenology-first approach over competing experimentally-driven approaches.

1. The rationale for the phenomenology-first approach.

Within the contemporary landscape of consciousness science, the majority of theoretical approaches gather fundamental data for theory-building from the observation of behavioural outcomes, such as subjective report, in conjunction with brain imaging data (Chalmers, 2000). These approaches hold that without neural and behavioural data we cannot possibly do meaningful science of consciousness, since without neural and behavioural data we would lack the objective observables at the basis of any scientific investigation (Doerig, Schurger, Hess, & Herzog, 2019; Herzog, Schurger, & Doerig, 2022). The starting point of consciousness science must thus coincide with experiments that pick out correspondences between conscious states and objectively (neural-behavioural) states.

Some of the most influential theories of consciousness committed to the this type of approach are the Global Neuronal Workspace Theory (GNWT) (Mashour, Roelfsema, Changeux, & Dehaene, 2020), Higher-order theories (Brown, Lau, & LeDoux, 2019), Attention Schema Theory (Graziano & Webb, 2015), and Recurrent Processing Theory (RPT) (Lamme, 2006, 2010). These theories gather data mainly from experiments where a significant correlation is found between specific subjective reports, or other behavioural outcomes, and specific neural properties. This observation is explained by the hypothesis that a particular neural property underpins the conscious state the subject is reporting.

In the context of experiment-driven approaches, the main problem is to understand the exact relation between reportability (or at least cognitive accessibility) and phenomenal consciousness, since some research programmes require a distinction between the

mechanisms underpinning these two phenomena (Block, 1995, 2011; Lamme, 2018), whereas others hold that studying consciousness independently of any behavioural outcome or report is simply outside the scope of science (Cohen & Dennett, 2011; Doerig et al., 2019; Naccache & Dehaene, 2007; for a comprehensive discussion, see (Kleiner & Hoel, 2021)).

Thus, within experiment-driven approaches, the debate seems mostly focused on i) whether data about cognitive reportability should be separated from data about phenomenal consciousness, and ii) if so, how to do it. Experiment-driven consciousness science, then, seems to have two issues. First, there is a theoretical problem: there is no consensus on which *concept* of consciousness should be used in consciousness research, since it is not clear whether phenomenal consciousness is identical with cognitive access or not, nor whether this distinction is useful at all (Block, 1995; Irvine, 2017; Overgaard & Grünbaum, 2012). Second, there is a methodological problem: since there is no consensus on how to conceptualize the *explanandum* phenomenon (the phenomenon that needs to be explained), there is no well-established methodology for studying consciousness scientifically (Irvine, 2012), and how to set it apart from all possible confounds (for a discussion, see (Michel & Morales, 2020))[1].

At the present stage, theories of consciousness based on the experiment-first approach seems to be systematically underdetermined (Michel, 2019), namely, the same data can be accounted for by different models with equal explanatory and predictive power: the claim that a certain model is best can be countered by proponents of a different model, who operate under different theoretical and methodological assumptions. This situation seems to lead to an impasse, which can be summarized in this way: at the present stage, theories employing an

[1] No-report paradigms, namely experimental paradigms that seek to track conscious states through physiological measures like eye movements or pupil dilation (Block, 2019; Tsuchiya, Wilke, Frässle, & Lamme, 2015), might help with this task, but see (Overgaard & Fazekas, 2016) and (Michel & Morales, 2020) for critical discussions.

experiment-driven approach do not have a well-defined methodology for deciding which model of consciousness fits best with currently available data.

This is acknowledged by several consciousness scholars. Victor Lamme states that "the debate [between theories that focus on phenomenal consciousness and those that focus on conscious access] seems to reach a stalemate" (Lamme, 2018, p. 1), whereas Phillips (2018) argues that "not only do we not know whether consciousness requires cognition, we do not know how to find out" (Phillips, 2018, p. 7).

Perhaps this impasse can be overcome from within the experiment-driven approach: finessing consciousness detection procedures and calibrating different measures of consciousness (Michel, 2021; Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008), paired with accumulation of new evidence, might help decide between competing models of consciousness.

In this paper, however, I want to explore the prospects and limitations of a different strategy. This is the strategy of claiming that, if we want to advance consciousness science, we must approach consciousness from a different angle, and this angle is provided by phenomenology itself. That is, in order to explain consciousness, we must start from consciousness itself. This is the "phenomenology-first" approach taken by IIT.

IIT is by no means the only research programme that starts with first-person data (e.g. (Rudrauf et al., 2017; Varela, 1996; Williford, Bennequin, Friston, & Rudrauf, 2018); for a discussion, see (Chalmers, 2004)), but its centrality in current neuroscientific debates on the neural basis of consciousness and its clear and well-developed theoretical structure place IIT at the forefront of phenomenology-first approaches. For these reasons, I will focus on IIT's specific version of developing the phenomenology-first approach.

According to IIT's phenomenology-first approach, although we need neural and behavioural data to *test* and *validate* theories of consciousness (IIT included), if we want to *build* one we need to extract the essential structure of experience from our own phenomenology. Phenomenologically gathered data are observations that are not publicly available and objectively measurable as traditionally conceived in science. Rather, they constitute a different kind of evidence, since they are observations of *how things appear to the subject*. A phenomenology-first approach to consciousness science can claim that the features that different appearances have in common just constitute phenomenal (from the Greek *phainomenon*, namely "that which appears") consciousness. In this sense, according to the phenomenology-first approach, phenomenologically gathered data, or first-person data, can be used scientifically if we acknowledge that the science of consciousness is the science of first-person appearances, and therefore how things appear to the subject must constrain and guide our understanding of consciousness itself[2].

This approach promises to resolve the theoretical and methodological issues that afflict experiment-driven approaches. First, it posits that phenomenal consciousness is the appropriate *explanandum* of consciousness science because it is manifest as an observational datum. Second, the phenomenology-first approach does not require an account of the relation between consciousness and reportability, since phenomenological observations are not data *for an extrinsic observer*. Rather, they are data *for the same subject* who is experiencing them. If the phenomenological observation of consciousness is *direct*, and does not involve neural and behavioural data, there is no risk of bringing confounding factors such as post-perceptual processes into the fundamental dataset upon which we build our theory of consciousness.

[2] This approach does not deny the importance of standard scientific data, which are objectively and publicly observable. Rather, it questions their epistemic primacy in the particular field of consciousness science, where the target phenomenon is considered subjective and private.

To clarify, according to IIT's phenomenology-first approach, phenomenological data are not subjective reports *about* consciousness, nor judgments or beliefs about one's own consciousness. It is not that there are two things: consciousness and the observation of consciousness. Rather, IIT seems to work with an account of introspection which is akin to the *acquaintance view* of introspection (Gertler, 2012; Russell, 1912): the idea is that we are in direct contact with our conscious states, and that there is no medium between how consciousness is and how it appears to the conscious subject. According to Russell's test, you are acquainted with something if you cannot possibly doubt about the object's existence. This seems in fact the (Cartesian) starting point of IIT, its zeroth postulate (Barbosa, Marshall, Albantakis, & Tononi, 2021, p. 2): consciousness exists and this much cannot be doubted[3].

If the acquaintance view of introspection is on the right track, in the case of consciousness, the distinction between appearance and reality vanishes (Kripke, 1980): the "real thing" is precisely phenomenology itself (i.e., how things *appear*)[4].

In this paper, I am not interested in assessing whether this is a convincing view of how introspection works (for criticisms of the acquaintance view see (Hill, 1991); see also (Dennett, 1991) and (Schwitzgebel, 2011) for criticisms against introspection in general, and (Spener, 2013) for a discussion). I will concede the point for the sake of the argument. Rather, my point here is that this view of how self-knowledge works seems to imply an important consequence for our explanatory practices in consciousness science: if there is no medium between consciousness and the subjective observation of consciousness itself, then

[3] Notice that consciousness, in IIT, means having a subjective perspective, and this seems to be shared even by Illusionists (Frankish, 2016; McQueen, 2019a).

[4] (Goff, 2017) and (Goff, 2020) take the acquaintance view of introspection as implying that we can grasp the essential *metaphysical* nature of consciousness. This is the Revelation argument. Proponents of the phenomenology-first approach for the science of consciousness do not need to endorse (and in fact they do not seem to endorse) this view on the metaphysical nature of consciousness, and can claim that the acquaintance view of introspection does not entail Revelation (see, e.g. (Rudrauf et al., 2017)). I thank an anonymous reviewer for pointing this out.

the datum (i.e., the observation) and the explanandum phenomenon (i.e., consciousness) coincide.

Thus, IIT's approach seems to reject, in consciousness science, the distinction between *data* and *phenomena* (Bogen & Woodward, 1988, p. 305) that the traditional, experiment-driven, approach implies. This is because in consciousness science the observer and the observed subject matter coincide: the phenomenon to be explained is precisely the immediate and direct experience the observer is having.

IIT's approach might have the advantage of not requiring an account of the relation between consciousness and reportability, but is founded on a relation between data and phenomena that is scientifically unorthodox, and it is not at all clear that a scientific explanation of consciousness can be provided by starting from phenomenological data (Herzog et al., 2022).

It is thus important to explore whether IIT's phenomenology approach can in fact use phenomenologically gathered data in order to provide a convincing explanation of consciousness. In order to do this, I will adopt a conditional strategy: I will thus accept for the sake of the argument that (i) IIT's phenomenology-first approach can be used to guide scientific enquiry on consciousness; (ii) IIT's foundation is in fact able to extract all the essential features of phenomenal consciousness, and (iii) that these features can be singled out via acquaintance with phenomenality itself[5]. I will argue, however, that even if we accept these foundational aspects of IIT, a problem for IIT as a scientific explanation of consciousness still remains (i.e., what I call "the self-evidencing" problem).

[5] Point (i) can be denied by maintaining that phenomenological data are no data at all, since scientific data are by definition third-person data derived from objective measurements (see, e.g. (Cohen & Dennett, 2011) and (Herzog et al., 2022). See also (Negro, 2020) and (Tsuchiya, Andrillon, & Haun, 2020) for replies). Point (ii) can be countered by denying that IIT's supposedly essential features of conscious experiences are in fact essential (Bayne, 2018), or by claiming that they do not exhaust the list of essential properties of experience (Brown, 2017). Point (iii) can be countered by denying that we get to know our conscious minds via acquaintance (see, e.g. (Frankish, 2016)). In this paper, I assume the truth of points (i)-(iii) in order to isolate the core issue I want to focus on, namely the logical structure of a scientific explanation of consciousness based on phenomenological data.

In particular, in order to evaluate whether IIT's methodology can truly be superior to the experiment-driven approach, we need to look at the particular structure of IIT first, and determine (i) whether IIT's phenomenology-first approach is able to "bootstrap" an explanation of consciousness from consciousness itself (Ellia et al., 2021, p. 10); and (ii) whether such an explanation avoids the impasse faced by the experiment-driven approach. So, let us examine IIT's structure first.

2. Building on Phenomenology

IIT's fundamental phenomenological building blocks are its *axioms,* which are supposed to capture the essential features of experience. These are: (a) intrinsic existence: consciousness exists from the intrinsic perspective; (b) composition: consciousness is structured, composed of phenomenological distinctions; (c) information: consciousness is specific, as each experience has its particular shape; (d) integration: consciousness is unified, as each experience is indivisible; (e) exclusion: consciousness is definite, as it has definite borders in space and time (for a complete description of the axioms, see (Oizumi et al., 2014) and (Tononi & Koch, 2015)).

These fundamental features of experience are used as a basis to derive a further set of assumptions, IIT's *postulates,* that are supposed to explain how the physical world must be in order to underpin the structure of experience picked out by the axioms. Thus, IIT's postulates set the stage to the *explanans,* namely the (physical) phenomenon that is supposed to explain the (phenomenal) phenomenon that needs to be explained. Each postulate is supported by the corresponding axiom, so we have: (a') intrinsic existence: a system must have causal power upon itself; (b') composition: a system must be structured; (c') information: a system in a state must be specific by constraining in a specific way its past (cause information) and future (effect information) states; (d') integration: a system must generate information that is

irreducible to that generated by its components; (e') exclusion: a system must generate only one irreducible cause-effect repertoire.

We thus have phenomenological building blocks (the axioms) and physical operationalizations (postulates) that seek to explain how these phenomenological features come to being. Given that the postulates are formulated in a language that is accessible to information theory, IIT expresses consciousness in information theoretic terms. IIT's central claim is that consciousness is explanatorily identical to integrated information[6], with Φ^{Max} being the measure of such information theoretic quantity, and therefore of consciousness too.

For Φ^{Max} to be a genuine explanation of consciousness, we need to make sure that (i) the axioms pick out all the essential features of consciousness, and thus provide the proper structure of the explanandum phenomenon; and (ii) the self-evident nature of the axioms is faithfully conveyed to the postulates. As mentioned earlier, in this paper, I will assume that the axioms do pick out all the essential features of consciousness (for a criticism, see (Bayne, 2018)), and I will instead focus on (ii).

The first question to address about the link between IIT's postulates and IIT's axioms concerns the type of inferential reasoning adopted to infer the former from the latter. As pointed out by Tononi (2017, p. 621), Grasso (2019, p. 51), and Koch (2019, p. 75), the inferential process through which postulates are extracted from the axioms seems to be an inference to the best explanation (IBE). That is, each postulate is supposed to *best explain*[7]

[6] I am using "integrated information" as shorthand for "maximally irreducible conceptual structure" (Oizumi et al., 2014, p. 14).

[7] Ellia et al. (2021) argue that the relation between introspective data and neuroscientific explanations is an inference to a "good enough" explanation. This might soften the requirement that the postulates must be true, but it does not undermine my overall point, since my argument is based on the *uniqueness* of the best explanation, and the "good enough" explanation provided by each postulate, in IIT, is still supposed to be unique.

why the particular feature of consciousness picked out by the corresponding axiom occurs, given the assumption that a physical world exists.

This type of inference is often thought to be *truth-conducive* (Psillos, 2002; see (Van Fraassen, 1980) for a criticism), and therefore, it seems well-suited to convey the self-evident truth of the axioms to the postulates, and, from there, to ground the accuracy of Φ^{Max} as a measure of consciousness.

The problem, however, is that this theoretical structure collapses if each postulate is not the *best* explanation of each corresponding axiom: if the set of chosen postulates is constituted by poor explanations, Φ^{Max} will be a poor measure of consciousness.

There is an interesting similarity between IIT phenomenology-first approach and the traditional methodology based on experimentally gathered neural and behavioural data: both methodologies can be conceived as trying to derive their explanatory models through a process of inference to the best explanation. As seen above, the problem for traditional approaches is that it does not seem possible to determine what counts as *best* between different theories, given the lack of consensus on a well-defined methodology for choosing a model over another. Traditional methodologies seem to be systematically underdetermined.

In the context of IIT, the underdetermination problem could apply if different explanatory hypotheses can be proposed to account for the same piece of phenomenological evidence[8].

Crucially, inference to the best explanation is a type of *contrastive* reasoning: the explanatory virtues of an explanation are (at least partially) identified by contrasting the explanatory power of that explanation with other candidate explanations (Lipton, 2017, p. 188). Given that postulates, in IIT, are derived from the axioms through an inference to the best

[8] I will use "explanatory hypotheses" and "operationalizations" interchangeably. Although the idea of explaining axioms might sound odd, the idea here is that, at least in the context of IIT, the postulates operationalize phenomenology by rendering it amenable to scientific investigation, and thus they are supposed to *explain* how the world must be in order to sustain phenomenology.

explanation, the logic fuelling the inference from axioms to postulates allows for multiple potential operationalizations of phenomenology. Therefore, we need to understand whether the chosen operationalization is the *best* possible explanation of the phenomenological evidence.

3. Bootstrapping an explanation of consciousness through self-evidencing explanations

According to IIT's phenomenology-first approach, one believes in the truth of explanatory hypotheses only in virtue of the truth of the evidence they are supposed to explain: the only evidence I have that a certain postulate is true is given by my own phenomenology. In cases like this, self-evidencing explanations might come in handy.

Self-evidencing explanations obtain every time "the information that the explanandum event has occurred […] constitutes, as we might say, an essential part of the only evidence available in support of that hypothesis" (Hempel, 1965, p. 372). These explanations are extremely common, and not only in scientific contexts. An every-day case of self-evidencing explanation involves the inference of the hypothesis "a burglar is afoot" from the observation of a footprint on the ground outside one's window (Lipton, 2004, p. 24). The observation of the footprint provides evidence for the hypothesis "a burglar is afoot", and in turn this hypothesis explains why the footprint occurred. Self-evidencing explanations have a circular aspect, but this circularity is benign: "an acceptable self-evidencing explanation benefits, as it were, by the wisdom of hindsight derived from the information that the explanandum event has occurred, but it does not misuse that information so as to produce a circular explanation" (Hempel, 1965, p. 373). Thus, a self-evidencing explanation occurs when an event provides evidence for the hypothesis that, if correct, would in turn explain the same observation that counts as evidence for that hypothesis. As Peter Lipton puts it: "Self-evidencing explanations exhibit a kind of circularity: H explains E while E justifies H. […] what is salient is that there

is nothing vicious here: self-evidencing explanations may be illuminating and well-supported." (Lipton, 2001b, p. 45).

In the case of IIT, a postulate would be a self-evidencing explanation of the corresponding axiom since the phenomenologically gathered axiom is an essential part of the evidence available in support of that postulate: we believe in the "goodness" of the postulate in virtue of the self-evident truth of the axiom. IIT can thus exploit self-evidencing explanations to "bootstrap" the explanation of consciousness from consciousness itself.

But, in order for this procedure to solve the version of the underdetermination problem for IIT, the self-evidencing explanation of consciousness must be (i) *acceptable*, namely, the self-evidencing must not turn vicious; and (ii) *successful*, namely, each self-evidencing postulate must be the best explanation of the corresponding axiom. Each axiom must then provide evidence for one, *and only one*, postulate, and this postulate must be the best explanation of the corresponding axiom.

If IIT wants to claim some explanatory superiority compared to traditional approaches based on neuroscientific and behavioural evidence, IIT needs to prove that the type of self-evidencing explanation its explanatory model (based on the axioms and postulates of the theory) implies, is in fact *successful*. Since traditional approaches fail to prove that a specific fit of neuroscientific and behavioural data into a model of consciousness is the *best* fit, IIT must show that its own inferential reasoning, based on IBE, does not suffer the same fate. That is, IIT must show that its own internal structure has some sort of explanatory superiority compared to the internal structure of theories that start from experimentally gathered data.

To restate, for the self-evidencing explanation of consciousness to be successful, in the context of IIT, it means that the circularity it implies should not be vicious (this is the *acceptability* criterion), and the mapping from the observation (i.e. each single axiom of IIT)

to the hypothesis (i.e. each single postulate of IIT) should pick out only *one hypothesis* – the best one (this is the *success* criterion). Let us examine whether IIT meets these two criteria for successful self-evidencing explanations.

4. The self-evidencing problem for IIT

In order for a self-evidencing explanation to be *acceptable*, the circularity of the explanation must not turn vicious. This happens when doubts about the occurrence of the *explanandum* event are dismissed by appealing to beliefs that depend on it. In the burglar example, if someone raises doubts about the fact that a burglar left a footprint outside of my window, I cannot use the belief "there is a burglar afoot" to dismiss their doubt, since I formed the burglar hypothesis in virtue of having observed the footprint. To dismiss doubts about the occurrence of the *explanandum* event, I need to appeal to considerations that are independent of the occurrence of the evidence (e.g. the high criminality rate in my neighborhood). These considerations are external to the self-evidencing circle, and constitute background knowledge that has been established prior to the observation of the footprint and my corresponding burglar hypothesis-formation. Appealing to such well-established background knowledge to dismiss doubts on the occurrence of the explanandum event makes the self-evidencing circle benign.

In this sense, IIT circularity does not seem to be vicious. Even if someone were to raise doubts about the instantiation of a certain phenomenological property[9] as essential property of experience (Bayne, 2018), IIT would not dismiss the doubt by appealing to its own explanatory model. For example, Bayne points out that the unity of consciousness, a phenomenological property picked out by the axiom of integration, might not be an essential

[9] Notice that in the case of IIT we need to move from an "event-based" talk to a "property instantiation-based" talk, but this does not affect my argument, as a nominalist about properties could formulate the argument in event-based terms without loss of content.

feature of consciousness, since the experience of patients with associative agnosia does not seem to be representationally unified (Bayne, 2018, p. 5). In IIT, the axiom of integration is explained by the integration postulate, which states that "the cause–effect structure specified by the system must be unified: it must be intrinsically *irreducible* to that specified by non-interdependent subsystems ($\Phi > 0$) across its weakest (unidirectional) link" (Tononi & Koch, 2015, p. 7). If IIT were to dismiss the doubt raised by Bayne by claiming that the experience of patients with dissociative agnosia is unified because their brain specifies an irreducible cause-effect structure, then the self-evidencing circle in the case of the inference from the integration axiom to the integration postulate would indeed be viciously circular.

But IIT could simply dismiss the doubt about the instantiation of the property by appealing to the *self-evident* nature of the axioms. That is, IIT could reply that axioms cannot be doubted by definition, as their truth is manifest in first-person experience, and that if doubts can be raised, it is just because the meaning of the axioms has been misunderstood – something has been lost in communicating the axiom.

My aim, here, is not to evaluate whether this move convincingly counters Bayne's challenge. Rather, what matters here is that this move does not make the self-evidencing circularity vicious, as the doubt about the observed property is dismissed through phenomenological considerations, and not through the hypothesis one forms in virtue of the observed property.

The self-evidencing circle in the case of IIT seems at least acceptable – i.e. it does not seem to turn vicious. But this is not enough for it to be successful, as the inferential reasoning IIT proponents use to formulate a postulate given the corresponding axiom (i.e., IBE) requires picking out only one postulate, and the self-evidencing dynamics does not *per se* guarantee that: a single piece of evidence can support different equally good hypotheses, and a self-evidencing circle will be generated for each of these competing hypotheses. If we want the

self-evidencing circle to be successful under IIT's version of inference to the best explanation we need a methodology for eliminating competing self-evidencing circles.

To see how this could be done, let us return to the burglar example. Imagine a competing hypothesis is advanced to explain the footprint: perhaps the footprint was left by a repairperson who was climbing the building to repair the roof. We now have two self-evidencing circles generated by the same piece of evidence, and we need to find a way to eliminate one of them. In cases like this, it is often thought that it is the well-established background knowledge that allows us to rank the competing explanations, and discard those that are deemed less explanatory: "an IBE-type of reasoning has a fine structure that is shaped, by and large, by the context. Explanations are, by and large, detailed stories. The background knowledge (or, beliefs) ranks the competitors. Other background assumptions determine the part of the logical space that we look for competitors" (Psillos, 2007, p. 443). In this case, I can appeal to considerations external to the circle, like my knowledge of high-criminality rate in the neighborhood and my knowledge that there is no repairment work scheduled for my building, to prefer the burglar hypothesis to the repairperson hypothesis.

We can now try to apply the same reasoning to IIT. Let us take, again, the integration axiom:

INTEGRATION: "consciousness is unified: each experience is *irreducible* to non-interdependent subsets of phenomenal distinctions" (Tononi & Koch, 2015, p. 7).

To restate, the corresponding postulate maintains that:

P_{IIT}: "the cause–effect structure specified by the system must be unified: it must be intrinsically *irreducible* to that specified by non-interdependent subsystems ($\Phi > 0$) across its weakest (unidirectional) link" (Tononi & Koch, 2015, p. 7).

But an alternative postulate, that would *prima facie* be supported by the integration axiom is conceivable, and it could be formulated in the following way:

P_{Alt}: A mechanism can contribute to consciousness only if the information it generates is broadcast into a global module that spatiotemporally binds the information coming from different sub-modules.

For the integration postulate to be a better explanation of the integration axiom than this alternative postulate, IIT proponents need to show *why* we have good reasons to eliminate the alternative postulate. Similarly, Kelvin McQueen (2019b, p. 88) shows that it is possible to formulate an information postulate that is different from that formulated by IIT, but fully compatible with the information axiom[10], while Merker et al. (2021) suggest that the intrinsicality axiom can be best accounted for by the notion of "point of view" introduced by the Projective Consciousness Model (Rudrauf et al., 2017). This seems to show that there is the possibility to conceive and formulate alternative postulates that would explanatorily compete with the postulates chosen by IIT, and this is a systematic problem for each single axiom-to-postulate inference. So, the reasoning I have applied here to the axiom of integration could in principle apply to every axiom (and corresponding postulate) of IIT. The problem arises the moment we realize that it is not clear on which *background knowledge* we could rely, in order to eliminate the alternative explanatory hypothesis. That is, the observation constituting the axiom (e.g. phenomenal unity) could be best explained by the corresponding postulate because the IIT postulate coheres with well-established background knowledge better than the alternative postulate, in the same way as the burglar hypothesis is better than the repairperson hypothesis because it coheres better with my knowledge of criminality rate in the neighborhood, repairment work schedule, and so on[11].

[10] McQueen draws a different conclusion than I do. This will be clarified in Section 5.

[11] Here, I follow Lipton (2001a) in taking explanatory considerations as a guide for inference. For the sake of the argument, I assume that explanatory virtues of good explanations are unity, coherence, ad integration with background knowledge (for a discussion on the compatibility between IBE and models of scientific explanation, see (Prasetya, 2021)).

But since IIT does not provide us with such relevant background knowledge, we are not in a position to evaluate which postulate is best. We can call this problem the "self-evidencing problem" for IIT, since it shows why the axioms of IIT do not support a unique formulation of the corresponding postulate, and without a unique mapping from axiom to postulate, the self-evidencing circularity the phenomenology-first approach is trying to exploit to explain consciousness from consciousness itself, might be at best acceptable, but not successful. In IIT, the axioms might be self-evident, but the postulates do not seem to be self-evidencing.

The self-evidencing problem is for IIT what the underdetermination problem is for experiment-driven theories of consciousness: the problem of finding out which explanatory model fits best with the available evidence. The peculiarity of the self-evidencing problem, however, is that it applies to approaches that take the explanation of consciousness to derive from within consciousness itself, and thus try to bootstrap consciousness out of phenomenological evidence via self-evidencing explanations.

5. Clarifying the specificity of the self-evidencing problem

The self-evidencing problem for IIT is also similar, in some respects, to the non-uniqueness problem for IIT. This is the problem of finding a unique and non-arbitrary path that goes from phenomenology to Φ (Hanson & Walker, 2021; Kleiner & Hoel, 2021, p. 11). The problem is that, if there is more than one way to map phenomenology into Φ, then the explanatory power of Φ *as a measure of consciousness* is undermined, as Φ is supposed to measure consciousness exactly because it derives from consciousness itself. And, therefore, it should not depend upon arbitrary choices posited by the theorizer (Barbosa et al., 2021; Merker, Williford, & Rudrauf, 2021).

The non-uniqueness problem for IIT can be formulated in several different ways: even if one accepts the set of axioms posited by IIT, it can be maintained that the formalization in

information-theoretic terms of the Φ measure can take several different forms (Barrett & Mediano, 2019; Hanson & Walker, 2021), or that the inference from axioms to postulates is invalid, because several alternative postulates can potentially be formulated for each axiom (McQueen, 2019b).

Here, I highlight how the self-evidencing problem for IIT I raise is importantly different from these versions of the non-uniqueness problem. First, differently from what (Barrett & Mediano, 2019) and (Hanson & Walker, 2021) have showed, the self-evidencing problem does not focus on the formalization of Φ as a measure of consciousness. The self-evidencing problem focuses instead on how a property picked out by an axiom (e.g. "integration of consciousness") can be best explained by the property picked out by the corresponding postulate (e.g. "irreducibility of cause-effect structure"). This is fundamentally a matter of logical reasoning, as it requires justifying the inference from axioms to postulates, rather than being a mathematical issue.

Second, contrary to what McQueen (2019b) argues, I do not claim that such inference is deductively invalid, simply because the inference from axioms to postulates is not a deduction. According to McQueen, who focuses on the axiom and postulate of information, the inference from the information axiom to the information postulate "is invalid because the axiom says nothing about past or future, cause or effect. These notions are simply smuggled into the information postulate" (McQueen, 2019b, p. 89). If the inference from axioms to postulates were a deductive inference, McQueen criticism would in fact be correct, since in deductive inference the conclusion cannot convey more information than that conveyed by the premises (D'Agostino & Floridi, 2009; Sequoiah-Grayson, 2008). But the inference from axioms to postulates is a form of IBE, and this form of reasoning makes it possible for the conclusion to convey more information than that conveyed by the premises: IBE is a form of *ampliative* reasoning (Psillos, 2002). The self-evidencing problem highlights that it is unclear

whether the axioms are faithfully conveyed to the postulates. And if each postulate chosen by IIT is not the best explanation of each corresponding axiom, then the self-evidencing circle between evidence and hypothesis cannot be successful. The uniqueness of the derivation of the Φ measure from phenomenology is thus threatened by the very logical reasoning fuelling the inference from IIT's axioms to IIT's postulates, as it is unclear how we can exclude different possible formulations of a postulate, even if we grant the self-evident truth of the corresponding axiom.

Finally, my argument also differs substantially from that of Merker et al. (2021), who claim that there is a mismatch between IIT's axioms and postulates, and therefore there does not seem to be any justification for the claim that IIT is a theory *of consciousness*. Rather, according to them, integrated information can be considered as a measure of network efficiency (for a reply, see (Tononi et al., 2022)) . My point, however, is not that the identity between consciousness and integrated information is mistaken, but that we do not have the instruments to tell whether the inference of IIT's postulates from IIT's axioms, which supposedly justifies IIT's central identity, is successful. This is because it is not clear that IIT's phenomenological ground is properly operationalized through a process of self-evidencing.

To be clear, I am not saying that IIT's postulates cannot be self-evidencing: the self-evidencing problem is not the problem that IIT's postulates *cannot* be best explanations of the corresponding axioms. Rather, the problem is that we do not have enough information to evaluate whether IIT's postulates are best explanations of the corresponding axioms, and, thus, whether they are *successfully* self-evidencing. This is because we are not given the background assumptions and background knowledge with which a given hypothesis should cohere. The self-evidencing problem for IIT is thus the problem of our inability to evaluate whether IIT's phenomenology-first approach can successfully explain consciousness from

self-evident features of consciousness itself, through a form of successful self-evidencing explanation. In order to solve the self-evidencing problem, IIT must make explicit the background knowledge and background assumptions that, together with the phenomenological observations provided by the axioms, support one formulation of a postulate over another.

6. A possible way forward?

As pointed out by Grasso (2019) and Tononi (2017), IIT seems to imply pandispositionalism, the ontological view according to which the only fundamental properties that exist are dispositional properties, whose existence is defined by their causal powers[12]. Thus, for IIT, the physical world is nothing but cause-effect powers. This ontological stance can count as background assumption that constrains the inference of a postulate from the corresponding axiom, since the postulate is supposed to tell us how the physical world should be, in order to underpin the phenomenological property picked out by the axiom. Coherently with pandispositionalism, and therefore with the idea that properties can be analysed in terms of causal powers, every IIT postulate is formulated in terms of cause-effect structure (Tononi & Koch, 2015, p.7). Pandispositionalism can importantly restrict the set of possible postulates from which we extract the best-explanatory postulate: for example, in the case of the integration axiom, the alternative postulate conceived in section 4 (P_{Alt}: A mechanism can contribute to consciousness only if the information it generates is broadcast into a global module that spatiotemporally binds the information coming from different sub-modules) can be excluded because it is not formulated in cause-effect terms. Given the unity of consciousness (integration axiom) and pandispositionalism, P_{IIT} is better than P_{Alt} because the

[12] It is not my intent, here, to explore the relation between the epistemology of IIT and its metaphysical assumptions and implications (for a discussion, see (Fallon & Blackmon, 2021)). I do not intend to suggest that pandispositionalism is uncontroversial either. What matters for the present purposes is that IIT proponents can resort to their endorsement of pandispositionalism to explain why the phenomenological evidence should be operationalized in the specific way proposed by IIT.

notion of irreducible cause-effect structure integrates better with pandispositionalism than the notion of information binding in a global module.

Note that the property of global information binding could perhaps be expressed, via reductive analysis, in terms of cause-effect powers, so to cohere with pandispositionalism. According to this reading, P_{Alt} would be excluded as the best explanation of the integration axiom not because incompatible with pandispositionalism, but because formulated at an inappropriate level of analysis: the best operationalization of the phenomenological evidence would be the one that analyses phenomenological properties in terms of cause-effect powers of the fundamental physical level, and not in terms of emergent levels of analysis (i.e., those of the special sciences).

This might be true, but for this move to work we need an argument to defend the claim that the right level of analysis, when operationalizing phenomenological properties, is the level of the cause-effect powers *of the physical*, rather than the cause-effect powers of an emergent level of analysis (e.g., the biological, or the psychological). IIT proponents, however, have yet to provide such an argument.

Moreover, although necessary to solve the self-evidencing problem, exhibiting background assumptions like pandispositionalism is likely to be insufficient. As pointed out by Psillos (2007), background assumptions can constrain the logical space from which explanations can be derived. In the case of IIT, pandispositionalism limits the logical space from which, given the axioms, postulates can be inferred. But different postulates can still be formulated from *within* this limited logical space, given that different postulates can be formulated in cause-effect terms, for each single axiom. For example, the integration axiom could be best-explained by the following postulate:

P_{Alt2}: "a mechanism contributes to consciousness if its cause-effect structure *fuses* the causal powers of its constituents".

This postulate is pandispositionalist in spirit, as it is formulated in cause-effect terms, and accounts for the unity of consciousness, as fusion is a physical operation that produces *unified* wholes (Humphreys, 1997). However, P_{Alt2} is also substantially different from the integration postulate formulated by IIT. And if IIT wants to exclude this alternative postulate, it must exhibit further assumptions and background knowledge, on top of pandispositionalism, according to which we can favour IIT's integration postulate to this newly formulated postulate.

Defining the physical world in terms of cause-effect structures, according to a pandispositionalist ontology, and requiring that our operationalization of the phenomenological evidence adhere with such ontology, certainly shapes and constrains the logical space in which we can look for the postulates that can best explain the axioms. But this is not enough, as several alternative postulates for each axiom can be conceived and formulated in terms of cause-effect powers: the self-evidencing problem remains.

Conclusion

In this paper, I have focused on IIT's phenomenology-first approach, and I have highlighted the way it integrates phenomenological evidence with the specific explanatory model IIT provides.

I have argued that IIT's approach differs from standard accounts in the neuroscience of consciousness, which normally start from experimentally-gathered evidence. At the present stage, experiment-driven approaches in consciousness science seem to have hit a roadblock related to difficulties in disentangling the mechanisms underpinning consciousness from those underpinning cognitive reportability. IIT can thus claim to be superior to experiment-

driven approaches because its peculiar phenomenology-first approach avoids this specific data-fitting issue.

I have suggested that IIT can exploit self-evidencing explanations to bootstrap the scientific explanation of consciousness from consciousness itself. Self-evidencing explanations are scientifically legitimate and their circularity can be virtuous, but for IIT to take advantage of this explanatory method, IIT must make sure that the circularity of self-evidencing explanations does not turn vicious, and that the inference from phenomenology to explanatory hypotheses picks out a unique explanation.

The main thesis of this paper is that IIT has a self-evidencing problem: this highlights that without exhibiting the background assumptions and background knowledge that underlie the inference from phenomenological evidence to explanatory hypotheses, the self-evidencing explanation of consciousness cannot be successful. The problem lies in our inability to evaluate whether the hypotheses chosen by IIT are in fact *best explanations* of the phenomenological evidence. Thus, IIT is affected by a data-fitting issue that is similar to that affecting experiment-driven approaches.

The scope of the self-evidencing problem is not to decide among IIT and other theories of consciousness. The scope, instead, concerns the internal structure of IIT; that is, how can we determine whether the theoretical foundation of IIT is well-suited to explain consciousness from phenomenology?

Nevertheless, solving the self-evidencing problem might have consequences with respect to the extrinsic relations IIT has with other theories of consciousness as well. Showing that the phenomenology-first approach can in fact exploit self-evidencing explanations to explain consciousness from consciousness itself would amount to showing that IIT is better suited to explain consciousness than approaches that start with experimental data, because it would put

IIT in a position of advantage with respect to the data-fitting issue that traditional approaches also face. IIT could justify why phenomenal consciousness (and not conscious access) is the real *explanandum* phenomenon of consciousness science, and it could provide a way to deal with it by exploiting a well-established scientific tool: self-evidencing explanations. However, while the self-evidencing problem is not solved, the explanatory power of IIT's phenomenology-first approach is on par with experimentally-driven approaches.

References

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy, 23*(3), 362. Retrieved from https://www.mdpi.com/1099-4300/23/3/362

Barrett, A. B., & Mediano, P. A. M. (2019). The Phi Measure of Integrated Information is not Well-Defined for General Phsyical Systems. *Journal of Consciousness Studies, 26*(1-2), 11-20.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci Conscious, 2018*(1), niy007. doi:10.1093/nc/niy007

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18(2)*, 227-247.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30*(5-6), 481-499. doi:10.1017/S0140525X07002786

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences, 15*(12), 567-575. doi:10.1016/j.tics.2011.11.001

Block, N. (2019). What Is Wrong with the No-Report Paradigm and How to Fix It. *Trends Cogn Sci, 23*(12), 1003-1013. doi:10.1016/j.tics.2019.10.001

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*(3), 303-352.

Brown, R. (2017). Integrated Information Theory is not a Theory of Consciousness. *https://onemorebrown.com/2017/08/05/integrated-information-theory-is-not-a-theory-of-consciousness/*. Retrieved from https://onemorebrown.com/2017/08/05/integrated-information-theory-is-not-a-theory-of-consciousness/

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences, 23*(9), 754-768. doi:10.1016/j.tics.2019.06.009

Chalmers, D. J. (2000). What is a neural correlate of consciousness? In T. Metzinger (Ed.), *Neural correlates of consciousness: Empirical and conceptual questions.* (pp. 17-39). Cambridge, MA, US: The MIT Press.

Chalmers, D. J. (2004). How can we construct a science of consciousness? In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences III* (pp. 1111--1119): MIT Press.

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences, 15*(8), 358-364. doi:https://doi.org/10.1016/j.tics.2011.06.008

D'Agostino, M., & Floridi, L. (2009). The enduring scandal of deduction. *Synthese, 167*(2), 271-315. doi:10.1007/s11229-008-9409-4

Dennett, D. C. (1991). *Consciousness Explained*: Penguin Books.

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition, 72*, 49-59. doi:https://doi.org/10.1016/j.concog.2019.04.002

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., P. Lang, J., . . . Tononi, G. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness, 2021*(2). doi:10.1093/nc/niab032

Fallon, F., & Blackmon, J. C. (2021). IIT's Scientific Counter-Revolution: A Neuroscientific Theory's Physical and Metaphysical Implications. *Entropy, 23*(8), 942. Retrieved from https://www.mdpi.com/1099-4300/23/8/942

Fink, S. B. (2016). A Deeper Look at the "Neural Correlate of Consciousness". *Frontiers in Psychology, 7*(1044). doi:10.3389/fpsyg.2016.01044

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies, 23*(11-12), 11-39. Retrieved from https://www.ingentaconnect.com/content/imp/jcs/2016/00000023/f0020011/art00002

Gertler, B. (2012). Renewed Acquaintance. In D. Smithies & D. Stoljar (Eds.), *Introspection and Consciousness* (pp. 89-123): Oxford University Press.

Goff, P. (2017). *Consciousness and Fundamental Reality*: Oup Usa.

Goff, P. (2020). Revelation, Consciousness+ and the Phenomenal Powers View. *Topoi, 39*(5), 1089-1092. doi:10.1007/s11245-018-9594-9

Grasso, M. (2019). IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience. *Journal of Consciousness Studies, 26*(1-2), 48-75. Retrieved from https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020001/art00004

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology, 6*(500). doi:10.3389/fpsyg.2015.00500

Hanson, J. R., & Walker, S. I. (2021). On the Non-uniqueness Problem in Integrated Information Theory. *bioRxiv*, 2021.2004.2007.438793. doi:10.1101/2021.04.07.438793

Haun, A., & Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy, 21*(12), 1160. Retrieved from https://www.mdpi.com/1099-4300/21/12/1160

Herzog, M. H., Schurger, A., & Doerig, A. (2021). First-person experience cannot rescue causal structure theories from the unfolding argument. doi: https://doi.org/10.31234/osf.io/s8a7n

Herzog, M. H., Schurger, A., & Doerig, A. (2022). First-person experience cannot rescue causal structure theories from the unfolding argument. *Conscious Cogn, 98*, 103261. doi:10.1016/j.concog.2021.103261

Hill, C. S. (1991). *Sensations: A Defense of Type Materialism*: Cambridge University Press.

Hohwy, J., & Frith, C. D. (2004). Can neuroscience explain consciousness? *Journal of Consciousness Studies, 11*(7-8), 180-198.

Humphreys, P. (1997). How Properties Emerge. *Philosophy of Science, 64*(1), 1-17. doi:10.1086/392533

Irvine, E. (2012). *Consciousness as a scientific concept: a philosophy of science perspective*: Springer.

Irvine, E. (2017). Explaining What? *Topoi, 36*(1), 95-106. doi:http://dx.doi.org/10.1007/s11245-014-9273-4

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness, 2021*(1). doi:10.1093/nc/niab001

Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*. Cambridge, MA: MIT Press.

Kripke, S. A. (1980). *Naming and Necessity: Lectures Given to the Princeton University Philosophy Colloquium* (Vol. 217): Harvard University Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494-501. doi:10.1016/j.tics.2006.09.001

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience, 1*(3), 204-220. doi:10.1080/17588921003731586

Lamme, V. A. F. (2018). Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), 20170344. doi:doi:10.1098/rstb.2017.0344

Lipton, P. (2001a). Is Explanation a Guide to Inference? A Reply to Wesley C. Salmon. In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical Approaches and Applications* (pp. 93-120). Dordrecht: Springer Netherlands.

Lipton, P. (2001b). What Good is an Explanation? In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical Approaches and Applications* (pp. 43-59). Dordrecht: Springer Netherlands.

Lipton, P. (2004). *Inference to the Best Explanation*: Routledge.

Lipton, P. (2017). Inference to the Best Explanation. In *A Companion to the Philosophy of Science* (pp. 184-193).

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron, 105*(5), 776-798. doi:10.1016/j.neuron.2020.01.026

McQueen, K. J. (2019a). Illusionist Integrated Information Theory. *Journal of Consciousness Studies, 26*(5-6), 141-169. Retrieved from https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020005/art00006

McQueen, K. J. (2019b). Interpretation- Neutral Integrated Information Theory. *Journal of Consciousness Studies, 26*(1-2), 76-106. Retrieved from https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020001/art00005

Merker, B., Williford, K., & Rudrauf, D. (2021). The Integrated Information Theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences*, 1-72. doi:10.1017/S0140525X21000881

Michel, M. (2019). Consciousness Science Underdetermined: A short history of endless debates. *Ergo: An Open Access Journal of Philosophy, 6*.

Michel, M. (2021). Calibration in Consciousness Science. *Erkenntnis*. doi:10.1007/s10670-021-00383-z

Michel, M., & Morales, J. (2020). Minority reports: Consciousness and the prefrontal cortex. *Mind & Language, 35*(4), 493-513. doi:https://doi.org/10.1111/mila.12264

Naccache, L., & Dehaene, S. (2007). Reportability and illusions of phenomenality in the light of the global neuronal workspace model. *Behavioral and Brain Sciences, 30*(5-6), 518-520. doi:10.1017/S0140525X07002993

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences, 19*(5), 979-996. doi:10.1007/s11097-020-09681-3

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol, 10*(5), e1003588. doi:10.1371/journal.pcbi.1003588

Overgaard, M., & Fazekas, P. (2016). Can No-Report Paradigms Extract True Correlates of Consciousness? *Trends in Cognitive Sciences, 20*(4), 241-242. doi:10.1016/j.tics.2016.01.004

Overgaard, M., & Grünbaum, T. (2012). Cognitive and non-cognitive conceptions of consciousness. *Trends in Cognitive Sciences, 16*(3), 137. doi:https://doi.org/10.1016/j.tics.2011.12.006

Phillips, I. (2018). The methodological puzzle of phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), 20170347. doi:doi:10.1098/rstb.2017.0347

Prasetya, Y. (2021). Which Models of Scientific Explanation are (In)Compatible with IBE? *The British Journal for the Philosophy of Science, 0*(ja), null. doi:10.1086/715203

Psillos, S. (2002). Simply the best: A case for abduction. In *Computational Logic: Logic Programming and Beyond : Essays in Honour of Robert A. Kowalski, Part Ii* (Vol. 2408, pp. 83-93): Springer Berlin.

Psillos, S. (2007). The Fine Structure of Inference to the Best Explanation. *Philosophy and Phenomenological Research, 74*(2), 441-448. doi:https://doi.org/10.1111/j.1933-1592.2007.00030.x

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology, 428*, 106-131. doi:https://doi.org/10.1016/j.jtbi.2017.05.032

Russell, B. (1912). *The Problems of Philosophy*: Home University Library.

Schwitzgebel, E. (2011). *Perplexities of Consciousness*: Bradford.

Sequoiah-Grayson, S. (2008). The scandal of deduction: Hintikka on the Information Yield of Deductive Inferences. *Journal of Philosophical Logic, 37*(1), 67-94. Retrieved from http://www.jstor.org/stable/41217823

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences, 12*(8), 314-321. doi:10.1016/j.tics.2008.04.008

Spener, M. (2013). Moderate scepticism about introspection. *Philosophical Studies, 165*(3), 1187-1194.

Tononi, G. (2017). Integrated Information Theory of Consciousness. In *The Blackwell Companion to Consciousness* (pp. 621-633).

Tononi, G., Boly, M., Grasso, M., Hendren, J., Juel, B. E., Mayner, W. G. P., . . . Koch, C. (2022). IIT, half masked and half disfigured. *Behavioral and Brain Sciences, 45*, e60. doi:10.1017/S0140525X21001990

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci, 17*(7), 450-461. doi:10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philos Trans R Soc Lond B Biol Sci, 370*(1668). doi:10.1098/rstb.2014.0167

Tsuchiya, N., Andrillon, T., & Haun, A. (2020). A reply to “the unfolding argument”: Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition, 79*, 102877. doi:https://doi.org/10.1016/j.concog.2020.102877

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences, 19*(12), 757-770. doi:10.1016/j.tics.2015.10.002

Van Fraassen, B. C. (1980). *The Scientific Image*: Oxford University Press.

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330-349.

Williford, K., Bennequin, D., Friston, K., & Rudrauf, D. (2018). The Projective Consciousness Model and Phenomenal Selfhood. *Frontiers in Psychology, 9*. doi:10.3389/fpsyg.2018.02571