Consciousness and Content from the Perspective of the Integrated Information Theory

Niccolò Negro

School of Psychological Sciences, Tel Aviv University

Abstract

This paper contributes to the debate about the nature of mental content from the perspective of the neuroscience of consciousness. In particular, I consider how one of the most influential neuroscientific theories of consciousness, the integrated information theory (IIT), understands the relation between consciousness and content. I conclude that it implies a form of *phenomenal intentionality theory* (PIT), the view that consciousness explanatorily grounds content, and for this reason proponents of PIT could find in IIT a neuroscientific ally. My main conclusion is that a higher degree of confidence in IIT should accompany a higher degree of confidence in PIT. In section 2, I show that major neuroscientific theories of consciousness implicitly commit to representationalism, the view that content explanatorily grounds consciousness. In section 3, I briefly present IIT, so to give the reader the necessary tools to understand the mechanics of my argument. In section 4, I argue that IIT implies a version of PIT, and that its theoretical apparatus could be re-interpreted to formulate a theory of content. In section 5, I argue that IIT is a form of PIT, which I will call '*structuralist PIT*'. In Section 6, I claim that IIT has the resources to push against the objection that nonconscious representations falsify PIT in general. I conclude that in virtue of this, proponents of PIT and IIT could work together to develop a more refined version of IIT as a theory of content, because if IIT turns out to be the correct theory of consciousness, this would help PIT too.

Keywords: Consciousness, Mental content, Integrated information theory, Phenomenal intentionality theory, Structuralist phenomenal intentionality theory.

1. Introduction

Consciousness and mental content are the two main protagonists of classical and contemporary debates in philosophy of mind.

An important question that has troubled philosophers for centuries is how to precisely understand the relation between consciousness and mental content, and in this regard we can broadly divide the conceptual space into three camps: I)

Argumenta 9, 2 (2024): 389—407
ISSN 2465-2334

DOI 10.14275/2465-2334/20230.neg
First published: 18 December 2023

separatism, namely the view that consciousness and content are totally different and independent phenomena; II) *representationalism*, namely the view that consciousness depends on content, and therefore an explanation of phenomenal properties necessarily depends on content properties; and III) *phenomenal intentionality theory* (PIT), namely the view that consciousness grounds mental content, and therefore an explanation of content properties necessarily depends on phenomenal properties.[1]

Despite separatism being traditionally the standard view on the matter (Kim 1998), this stance has been recently challenged either by representationalism or by PIT (Horgan and Tienson 2002; Mendelovici and Bourget 2020; Pautz 2020—but see Márton 2022).

This paper contributes to this debate by providing an interpretation of how one of the leading theories in the neuroscience of consciousness, the integrated information theory (IIT) (Oizumi, Albantakis, and Tononi 2014; Tononi 2004; Tononi, Boly, Massimini, and Koch 2016), frames the relation between consciousness and content. This is relevant because I will argue that IIT supports PIT, and provides a neuroscience-based account to solve the traditional challenge that non-conscious states seem to pose to that view.

The scope of this paper is not to claim that PIT is correct, but to support instead the conditional statement that *if* IIT turns out to be the best available theory of consciousness, *then* it would provide indirect support for believing in PIT. Nonetheless, this is in itself a relevant result, because it gives reasons to the proponents of PIT to side with IIT proponents, and work towards a more refined integration of PIT with IIT in order to derive a neuroscience-based version of PIT.

My argument is the following:

P1) If separatism is false, then either representationalism is true or PIT is true.

P2) Major neuroscientific theories of consciousness assume representationalism.

P3) IIT is the only major neuroscientific theory of consciousness that implies PIT.

P4) IIT helps solve important challenges for PIT.

Therefore,

C) If IIT is better than other major neuroscientific theories of consciousness, then we have good reasons to claim that PIT is better than representationalism.

Let me add some considerations to clarify the dynamics of the argument. P1 is intentionally in conditional form, as I do not intend to argue for the falsity of separatism and I will not provide extensive support for this premise. Moreover, although there might be interesting ways to combine representationalism and PIT (more on this in Section 4), I take the two views to be mutually exclusive because based on the *explanatory* priority granted to either consciousness or content, and I take explanations to be generally asymmetric. So, if it is true that content explains consciousness, it cannot be true that consciousness explains content (and *vice versa*).

I also take IIT and other major theories of consciousness as genuine competitors in virtue of being mutually exclusive explanations *of the same phenomenon*.

[1] I refer to PIT as a general and coarse-grained view, without presupposing any particular way in which the idea that consciousness grounds content could be made more precise and specified. For a discussion, see Bourget and Mendelovici 2016.

This assumption could be challenged by claiming that IIT and other theories of consciousness have different explanatory targets, and that they are not incompatible. This point is not devoid of merits, but an extensive treatment of this issue goes well beyond the scope of this paper. Here, I will treat IIT and other theories of consciousness as competing theories with the same explanatory target by adopting a descriptive stance, and by noticing that the main proponents of these theories do consider them as directed to the same phenomenon (Melloni et al. 2023; Melloni, Mudrik, Pitts, and Koch 2021).

Finally, the argument is supposed to support PIT, not IIT. I will not provide here an argument in support of IIT. Rather, the claim here is that IIT could be used to support PIT. A phenomenal intentionality theorist could still maintain that even if IIT is sufficient to support PIT, it is not a necessary condition for it, and therefore PIT could still be true even if IIT turns out to be false. As stated above, the modest goal of this paper is to claim that a higher degree of confidence in IIT should accompany a higher degree of confidence in PIT, and therefore proponents of PIT should take a favorable attitude towards IIT, and consider it as a potent ally.

Some terminological considerations for the sake of clarity: Unless otherwise stated, by 'consciousness' I mean phenomenal consciousness (Block 1995), namely the "what it is like" to be a subject (Nagel 1974). I will thus use the term 'phenomenal states' meaning mental states with the property of being phenomenally conscious.

I will also help myself to the notion of 'intentionality', namely the property of the mind to be directed towards an object, and by which mental states have *contents*, or *meaning*. I will use meaning and content interchangeably, describing *mental* meaning and content. The term 'intentional states' refers to mental states with intentional contents.

The label 'representationalism' for the view that intentional properties ground phenomenal properties is slightly inaccurate, because it assumes that the nature of intentionality is representational, but there are ways to explain intentionality in non-representational terms (Hutto and Myin 2013; for a discussion, see Schlicht 2018). However, the "intentionality-first" views I am going to discuss in this paper assume that intentionality is essentially representational, and therefore the label 'representationalism' is an effective shortcut for the present purposes.

This is the structure of the paper: In section 2, I will support P2 by showing the implicit commitment that major neuroscientific theories of consciousness have towards representationalism. In section 3, I briefly present IIT, so to give the reader the necessary tools to understand the mechanics of the argument. In section 4, I argue that IIT implies PIT, and that its theoretical apparatus could be reinterpreted to formulate a theory of content. In section 5, I complete the re-interpretation of IIT as a theory of content, and I argue that it could be seen as a form of PIT, which I will call '*structuralist PIT*'. Section 6 defends the idea that IIT has the resources to push against the objection that non-conscious representations posited by cognitive neuroscience falsify PIT in general. I conclude that in virtue of this, proponents of PIT and IIT could work together to develop a more refined version of IIT as a theory of content, because if IIT turns out to be the correct theory of consciousness, this would help PIT too.

2. Consciousness and Content in the Neuroscience of Consciousness

In the last thirty years, many neuroscientists have attempted to establish different research programs that aim to investigate the neural, biological, and mechanistic basis of consciousness. These efforts seek to provide a scientific understanding of consciousness, a phenomenon that has been a main protagonist of philosophical debates for centuries, and could be described as the most peculiar and characteristic feature of the mind (Searle 1992).

This surge in the scientific interest in consciousness has helped establish the science of consciousness as an interdisciplinary field that directly tackles many relevant questions on consciousness and its relation to neural structures and activity (Francken et al. 2022; Signorelli, Szczotka, and Prentner 2021; Yaron, Melloni, Pitts, and Mudrik 2022). However, debates within this exciting field rarely address questions about the other main protagonist of perennial debates in philosophy of mind, namely intentionality.

The lack of a systematic treatment, within the science of consciousness, of how consciousness and intentional content relate, could *prima facie* suggest a separatist approach, but I believe that the default attitude in the science of consciousness towards this problem should be seen as broadly representational. This is because the majority of scientific theories of consciousness are constructed around the problem of which kind of property makes a certain neural event conscious, and this neural event is typically interpreted as a representation of an external stimulus (for an antecedent philosophical account of this sort, see Tye 1995). In other words, these approaches take intentional states to be representational states, and phenomenal properties (if they are taken to exist at all) are explained in virtue of the physico-functional properties of those representational states.

For example, the Global Neuronal Workspace Theory (GNWT) (Mashour, Roelfsema, Changeux, and Dehaene 2020) claims that a representation becomes conscious when it is broadcast into, and becomes available to, a global workspace of various consumer systems; Higher Order Theories (HOT) (Brown, Lau, and LeDoux 2019) claim that a representation becomes conscious when it becomes the target of a higher-order representation; the Recurrent Processing Theory (RPT) (Lamme 2006) claims that a representation is conscious when it enters a recurrent information processing realized by feedback, *i.e.*, re-entrant, connections (for a review, see Seth and Bayne 2022).

These theories of consciousness do not specify the type of relation that makes a neural representation have the content that it has, but they can be aided in this task either by *tracking* theories of content, cashed out in a causal (Fodor 1990) or *teleosemantic* (Millikan 1989; Neander 2017; Shea 2018) form, or by *conceptual role* theories of content (Block 1998). The core idea of causal accounts of content is that the content of a mental representation is fixed by the worldly object the representation stands in causal relation with, while the core idea of teleosemantic approaches is that the content of a representation is fixed by how that representation is used by relevant consumer systems. On the other hand, functional/conceptual role semantic claims that the content of a representation is fixed by the inferential role it plays in an interconnected network (see Neander 2008 for a comprehensive discussion).

The marriage between neuroscientific theories of consciousness and these types of theories of content seems to be warranted by a materialistic approach that attempts to explain every aspect of the mind in physico-functional terms. That is, if one wants to explain consciousness in a materialist framework by employing representational tools, those very tools need to be firmly grounded on physico-functional terms, meaning that both the vehicle and the content of a representation should be materialism-friendly. This is what tracking theories and conceptual role theories of content seek to provide. Thus, theories of consciousness like GNWT, HOTs, and RPT could take advantage of materialistic theories of content to refine and make precise the relation between consciousness and content.

To be clear, there are major differences between the theories of consciousness I have considered. Above all, GNWT and HOTs are formulated as *cognitive* theories at the functional level,[2] whereas RPT is primarily formulated as a *neural* theory at the realiser level. What matters for the present purposes is that all these scientific accounts of consciousness individuate first a representation, and then they explain consciousness as a (functional or physical) property of that representation. The discussion so far thus supports P2, namely that most major neuroscientific theories of consciousness assume some version of representationalism.

This representation-first approach to consciousness is prominent and influential, but there is another leading theory of consciousness that does not seem to endorse this standpoint: The integrated information theory (IIT).[3] IIT is a non-cognitive theory of consciousness, in the sense that it treats consciousness as a fundamental entity, rather than a property exclusively of mental states, and is not straightforwardly interpretable as representational insofar as neural representations do not appear in its explanatory apparatus. Moreover, IIT's metaphysical implications seem to diverge from the traditional materialistic picture (Cea 2021; Grasso 2019; Negro 2022; Tononi and Koch 2015). The question, then, is whether IIT's peculiar account of consciousness has also peculiar implications for mental content. I argue that it does: There is a sense in which IIT suggests that mental content is grounded in consciousness, turning the traditional representation-first approach on its head.

3. IIT: A Brief Presentation

I will provide a brief exposition of IIT so to have the essential resources for my argument. The reader can find comprehensive presentations of IIT in (Oizumi et al. 2014; Tononi 2015; Tononi et al. 2016) and a summary of the latest version of the theory (called 'IIT 4.0') in (Albantakis et al. 2023).

IIT occupies a peculiar position in the science of consciousness, because its starting point is consciousness itself, not the brain. IIT defines consciousness as i) existing; ii) existing intrinsically; iii) being structured; iv) being specific; v) being unified; and vi) being spatiotemporally definite (Oizumi et al. 2014; Albantakis et al. 2023). From these phenomenological observations, which IIT calls "axioms" and considers self-evident and indubitable, IIT extracts a corresponding set of "postulates", which are theses that tell us how a physical system must be in order to implement the axioms. Postulates are therefore considered an operationaliza-

[2] This does not mean they do not provide sophisticated neurophysiological details of how the function is implemented.

[3] There are other theories of consciousness that are both cognitive and non-representational (*e.g.*, O'Regan 2014), but I am not going to focus on those here—for a comprehensive overview, see Seth and Bayne 2022.

tion of consciousness (the phenomenon picked out by the axioms), and are expressed through the language of cause-effect powers given that IIT assumes that the physical world is made of causal powers (Tononi, Albantakis, Boly, Cirelli, and Koch 2022).

Now, IIT formalizes these postulates through an information-theoretic measure (Barbosa, Marshall, Streipert, Albantakis, and Tononi 2020), that expresses the quantity of maximal integrated information a system specifies *for itself*, and thus *intrinsically*.[4] Information, here, should be interpreted as how *specific* the cause-effect power of the system is, since it measures how a system's components can potentially make and take a difference to and from the other components, while integration means *irreducibility*, since it amounts to how the whole makes a difference (to itself) that goes beyond the difference-making power of its components. The cause-effect powers that matter for consciousness, according to IIT, are also *compositional*, since different causal relations among components can make a difference to the whole in different ways, and *maximal*, since only the maximum of integrated information can define the boundaries of the physical substrate of consciousness from the intrinsic perspective (*i.e.*, without arbitrary decisions of an external observer). The quantity of maximally intrinsic, compositional, specific, and irreducible cause-effect powers of a system in a state is expressed by Φ^{Max}. Given that the formalism deriving Φ^{Max} is extracted from IIT's postulates, and IIT's postulates are extracted from the axioms, which are thought of as defining consciousness itself, IIT concludes that Φ^{Max} just is the measure of consciousness. In a nutshell, IIT posits an identity between consciousness and intrinsic integrated information.[5]

IIT thus arrives at an account of consciousness starting from consciousness itself, without mentioning brain functions or cognitive capacities. Again, consciousness as integrated information is a fundamental property, not a psychological property (Tononi and Koch 2015): *Any* physical[6] system that instantiates Φ^{Max} is a conscious system.

However, IIT does admit that brains are particularly well suited to specify Φ^{Max}, and that integrated information, despite being distributed across the physical world like mass and charge, peaks in biological brains (and perhaps neuromorphic artificial systems): Structures in these systems can constitute the neurobiological basis of consciousness, despite such a basis being only one of the multiple possible bases of consciousness in the Universe. In particular, IIT claims that the grid-like structure of the brain's posterior cortices is structurally apt to generate high levels of Φ^{Max}, and therefore consciousness (Boly et al. 2017; Grasso, Haun, and Tononi 2021).

This shows that IIT, despite not being a *neurobiological* theory of consciousness, has direct implications for the neurobiological basis of consciousness. In

[4] IIT 4.0 explicitly separates the axiom of existence from that of intrinsicality. On the one hand, existence is translated into the idea that to exist is to have cause-effect powers; on the other hand, intrinsicality establishes that cause-effect powers must be exerted from a system to *the system itself* (Albantakis et al. 2023: 5).

[5] Here, I will use 'integrated information' as a shortcut for "a maximally irreducible, specific, compositional, intrinsic *cause-effect structure*" (Haun and Tononi 2019: 7).

[6] As explained in Tononi et al. 2022 and Albantakis et al. 2023, physicalism is taken by IIT to be an operational view, namely a useful framework to explain, manipulate, and predict a phenomenon of interest, and not as the ontological view for which everything that exists is physical.

what follows, I will show that IIT, despite not being a *cognitive* theory of consciousness, has direct implications for the relation between consciousness and *mental content*.

4. Meaning as Integrated Information

The close association between consciousness and information predicated by IIT might suggest that IIT attempts to distil consciousness from a manipulation of symbols. But this is not quite right, since, according to IIT, intrinsic integrated information is essentially semantic, and the way the system informs itself corresponds to *meaning*. With Tononi's words:

> For the IIT, mechanisms generate meanings. Moreover, only the mechanisms within a single complex do so. [...] what is meaningful is each individual experience, and its meaning is completely and univocally specified by the shape of its quale (Tononi 2008: 238).

Let us unpack this point. IIT equates meaning to the content of an experience, which corresponds to the informational relationships between the various possible states the physical substrate of consciousness (*i.e.*, the 'complex', in IIT terms) can be found in. The standard objection that semantic content cannot be distilled from the information conveyed via syntactical relations (Searle 1980 2013) can be countered by maintaining that IIT's information is observer-independent, and therefore intrinsic to the system that instantiates it, whereas Shannon information is observer-dependent. For IIT, intrinsic information does not measure how much uncertainty is reduced by manipulating strings of symbols (as Shannon information does), but it measures, instead, the difference-making power of a physical state upon itself, and this intrinsic difference-making power fixes both consciousness itself and *the particular way* consciousness is—its content, or meaning (see Mindt 2021 for an insightful discussion).

To understand this point, we need to clarify a theoretical construct of IIT that will be particularly relevant for the present discussion. This is the notion of *cause-effect structure* (or Φ-structure): A cause-effect structure specifies the *shape* of the integrated information instantiated by the system. On the one hand, Φ^{Max} tells us *how much* consciousness a system has (*i.e.*, the quantity, or level, of consciousness); on the other hand, the cause-effect structure tells us *of what* the system is conscious (*i.e.*, its content): The account of conscious contents predicated by IIT thus maintains that each single experience corresponds to a specific shape of the informational structure generated by a physical system in a state, which represents how specifically the system's components constrain each other.

Such a structure can be geometrically represented by plotting it in a multidimensional space, called 'qualia-space', where each axis is given by a possible state of the system, and the coordinates are given by the probability distributions over all the possible states the system can find itself in, given the current state. Every content of consciousness, then, is identical to a specific cause-effect structure. Particular phenomenal properties like seeing red, or smelling coffee, correspond to φ^{Max}, namely the irreducible causal structures specified by *parts* of the complex, while the global state of consciousness (Bayne 2007; McKilliam 2020) the subject is experiencing here and now corresponds to the global constellation of $\varphi\text{s}^{\text{Max}}$.

The idea that a cause-effect structure fixes a particular content of consciousness, together with the claim that integrated information is essentially semantic, seems to support the idea that IIT has implications for a theory of content.

These implications can be made more precise by considering two theoretical results associated with the idea that contents of consciousness are fixed by cause-effect structures. The first theoretical result is a form of *phenomenal holism*. This is the idea that each experience constitutively depends on the relations it has with other experiences (Fink, Kob, and Lyre 2021; Lyre 2022). According to phenomenal holism, the redness of red, for example, feels the way it feels in virtue of its relations with the greenness of green, the blueness of blue, and so on.

A version of holism is implied by IIT insofar as experiences have the qualities they have in virtue of the informational relations instantiated by their physical substrates. This means that a global state of consciousness that includes experiences in different modalities (visual, gustatory, auditory, etc.) corresponds to a giant informational structure generated by parts of the cortex that are unified in virtue of integrating information as a single entity (*i.e.*, the complex), with states of, say, the visual cortex making a difference to states of, say, the auditory cortex (Balduzzi and Tononi 2009). As Tsuchiya puts it:

> The visualness of visual experience is determined not only by the way visual neurons interact with other visual neurons, but it also depends on how the visual neurons interact with auditory neurons and other neurons within the complex (Tsuchiya 2017: 7).

This does not necessarily correspond to the idea that the experience of, say, seeing red is partly constituted by the experience of hearing a middle C. Rather, the idea is that an experience is what it is in virtue of the form of the cause-effect structure, and such structure is essentially a relational and holistic entity.

The second implication IIT has for a theory of mental content is *internalism*: According to IIT, what really determines the contents of consciousness are the intrinsic informational relationships of the complex, and not its connections with input and output systems. These connections (in particular to input systems) might have a role in modulating the background conditions that indirectly influence the complex, but they are not constitutive of the content-fixing conditions. This is because the IIT formalism shows that one could severe the connections between the complex and input-output systems without any loss of integrated information, and therefore without affecting the integrated information structure that determines the particular content of the experience (Oizumi et al. 2014). This is nicely captured by this passage by Tononi:

> Consciousness *qua* integrated information is intrinsic and thus solipsistic. In principle, it could exist in and of itself, without requiring anything extrinsic to it, not even a function or purpose. For the IIT, as long as a system has the right internal architecture and forms a complex capable of discriminating a large number of internal states, it would be highly conscious. Such a system would not even need any contact with the external world, and it could be completely passive, watching its own states change without having to act (Tononi 2008: 239).

To sum up, IIT posits that meaning is co-instantiated with consciousness as integrated information, and that what fixes a particular meaning, or content, of an experience is the relational/informational profile of its physical substrate. This implies that contents of consciousness are holistic and internal, in the sense that input and output systems play no role in constituting it. I will now try to interpret this IIT picture in light of philosophical theories of content to bridge the gap between conscious content and mental content.

5. IIT as Structuralist PIT?

In section 1, I have argued that materialistic theories of content, which try to naturalize mental content by explaining it in physico-functional terms, could be combined with, and complement, major "representation-first" theories of consciousness, like GNWT, HOTs, and RPT. Here, I claim that a "consciousness-first" theory like IIT can instead be complemented by PIT, the view that consciousness is explanatorily prior to intentionality: If we want to understand why our mental states have the content that they have, we need to understand first why and how they feel the way they feel (Bourget and Mendelovici 2016; Kriegel 2013; Mendelovici 2018).

Versions of PIT can differ with respect to strength and to their relation with representationalism (see Bourget and Mendelovici 2016 for a comprehensive introduction). On the one hand, strong PIT claims that all intentional states are phenomenal states, while a more moderate version of PIT claims that intentional states are partly grounded on phenomenal states.

On the other hand, some versions of PIT, if they endorse the view that intentionality depends on consciousness without being identical to it, will be incompatible with representationalism, whereas versions of PIT that draw an identity between intentional properties and phenomenal properties (Mendelovici 2018) can be compatible with representationalism, if it turns out that phenomenal properties are essentially representational—here, the view would be that mental states are representational, but the content of that representation is given by the phenomenal properties of that mental state.

Given this brief summary, the question here is how this theory of content can complement IIT. The central idea is that, according to IIT, an integrated information structure *is* meaning, and therefore the structure itself is content-fixing. But the integrated information structure is extracted from consciousness itself (*i.e.*, IIT's axioms), and therefore it is first and foremost a theoretical construct posited to explain consciousness: In this sense, in IIT, consciousness is explanatorily prior to intentionality. Integrated information explains consciousness, and the shape integrated information takes fixes the meaning of a particular experience. Thus, once we have established that integrated information is explanatory with regard to consciousness, we can also use integrated information to explain content. In this sense, IIT's theoretical architecture seems to align nicely with the basic tenets of PIT. This much is also flagged by Mindt, who claims that "meaning is phenomenally constituted according to IIT, "the meaning is the feeling" (to borrow Giulio Tononi's phrasing for this)" (Mindt 2021: 12).

In the remainder of this section, I will address the questions of whether IIT could be seen as a strong or moderate version of PIT, and I will then show that IIT is not compatible with representation-friendly versions of PIT. In contrast, I will present IIT as version of phenomenal intentionality theory that I will call 'structuralist PIT'.

First, there are good reasons to think that IIT should be seen as *strong* PIT. This is because, as seen above, according to IIT, content co-occurs with consciousness *qua* integrated information. Given that IIT posits an explanatory identity between consciousness and integrated information, and integrated information *just is* meaning, the identity between consciousness and meaning, or content, seems to follow. Because of this identity, it seems that all intentional states are phenomenal states, insofar as they are integrated information states. IIT seems to suggest a strong reading of PIT.

Second, IIT does not seem to be compatible with a representational version of PIT, because according to IIT integrated information states are not representational. In order to unpack this point, we need to clarify the relation between integrated information and the external world.

In the previous section, I have argued that IIT endorses a form of internalism according to which the content-fixing conditions are entirely within a subject's head (more precisely, they are defined by the boundaries of a Φ^{Max}-generating physical system, which is probably to be found in the posterior brain areas).[7] However, IIT does provide an account of how the integrated information states of a Φ^{Max}-generating systems are indirectly connected to the external world: The notion of "cause-effect matching" enters IIT's picture for exactly this reason. From a formal point of view, Tononi defines matching as "a measure that assesses how well the integrated conceptual structure generated by an adapted complex fits the causal structure of the environment" (Tononi 2012: 306-307).[8] Indeed, IIT admits that in the context of biological organisms like us, the structure of consciousness *must* be somehow connected to the structure of the environment:

> Through natural selection, epigenesis, and learning, informational relationships in the world mold informational relationships within the main complex that "resonate" best on a commensurate spatial and temporal scale. [...] In this way, qualia—the shapes of experience—come to be molded, sculpted, and refined by the informational structure of events in the world (Tononi 2008: 240).

Thus, consciousness as integrated information *matches* the structure of the environment because the intrinsic integrated information constituting consciousness matches the extrinsic information between events in the environment under a selectionist perspective. This is the idea, derived partly from the work of Gerald Edelman (1987; for a review, see Seth and Baars 2005), that throughout development neuronal populations are selected in virtue of changes in synaptic strength that must adapt to changes in the environment.

The crucial point for the present discussion is that IIT does not understand this structural attunement between intrinsic and extrinsic informational structures, measured by cause-effect matching, to be *representational* in the standard sense that this term has in cognitive sciences, if the "job description" of a representation includes playing some sort of functional role (Facchin 2021; Gładziejewski 2015; Ramsey 2007; Shea 2018). This is because the informational

[7] In principle, this does not mean that these conditions *must* be "inside the skull"—if two cortices of two different subjects were physically connected in a way that generates a maximum of integrated information, the conscious system would comprise the two cortices of two different skulls.

[8] The term 'conceptual structure' was used in previous versions of the theory, and refers to what the current version calls 'cause-effect structure'.

content constituting the quale is not necessarily *used* by the system and therefore has no necessary functional profile.

Rather, the picture IIT implies is this: The environment exerts causal pressure on the brain, and the brain entertains states whose informational relationships match the environment's regularities.[9] But the informational structures generated by these brain states (cause-effect structures) are not themselves necessarily connected causally to the stimulus, nor they need to be exploited by the brain's consumer systems: Cause-effect structures are just the result of a structural attunement between brain and environment; an attunement directed by the principles of natural selection. Because of these evolutionary principles, the difference-making power of neurons in the brain happens to mirror the causal structure of things out in the world. This means that an external stimulus does not need to be robustly encoded in certain patterns of neuronal activity, since silent neurons (not *inactivated*) still retain their difference-making powers, and therefore can still contribute to the cause-effect structure. As a result, the mirroring between internal and external causal structures is not encoded in what neurons do, but in what they could potentially do. Matching, in itself, is a measure of this mirroring *in the here and now*, but high matching in a system results from evolutionary imperatives biological systems must comply with (Tononi, Sporns, and Edelman 1996).

Matching, thus, plays a role in determining the semantic aspect of our mental lives by connecting intrinsic causal structures with extrinsic causal structures in the environment (Albantakis, Hintze, Koch, Adami, and Tononi 2014). However, this is only a contingent role, and not a *constitutive* one: Mental contents are fixed by intrinsic cause-effect structures even if such structures have not undergone the contingent process of matching. This implies that for IIT the semantic aspect of the mind does not necessarily require a connection with the external world: The true "meaning" of a mental state corresponds to the intrinsic cause-effect structure specified by the system's intrinsic integrated information, and not to a worldly object (or our connection to it).[10]

This discussion of matching in IIT helps see that the relation IIT posits between integrated information states (*i.e.*, intrinsic cause-effect structures) and the external world is not a representational relation. Importantly, this also clarifies that the structural mirroring posited by IIT's matching is not conducive to structural representations in the sense adopted by, for example, Shea (2018) and Lee (2019), and is also different from the representational notion of matching in (Dalbey and Saad 2022). Therefore, IIT seems to be at odds with representation-friendly accounts of PIT.

A further step that helps place IIT in the landscape of theories of content is to assess its relation with conceptual role semantics (CRS) (Block 1998). According to CRS, the meaning of a mental token is fixed by the network of inferences it allows. One peculiar feature of this account of content is that it is holistic, in the

[9] A purely correspondence-based account of representation might consider this picture as "representational". See Baker, Lansdell, and Kording 2022 for a discussion.

[10] From a philosophical angle, it could be said that matching is a way to connect "meaning as sense" (or giveness) with "meaning as reference". However, given that, according to IIT, the connection between the intrinsic cause-effect structure and the external world is not content-fixing, "meaning as reference" is not genuine meaning. It could be said that cause-effect structures, at best, *quasi-refer* to worldly objects, since they are only indirectly influenced by them. I thank an anonymous reviewer for suggesting this interesting reading of IIT's matching.

sense that the content of a mental state is defined by the various relations it has with many, if not all, members of the network in which the state is embedded.

This aspect of CRS could be seen as a point of contact between IIT and CRS. As seen in Section 4, IIT seems to espouse a version of phenomenal holism for which the state of a neuron in an area of the cortex can influence the state of neurons in different areas. This means that, for IIT, as for CRS, contents are *essentially relational*, and an interesting research project for formalizing through category theory this relational nature of contents of consciousness in IIT is currently ongoing (Tsuchiya and Saigo 2021; Tsuchiya, Taguchi, and Saigo 2016; Zeleznikow-Johnston, Aizawa, Yamada, and Tsuchiya 2023).

This shows that there is in fact an interesting similarity between IIT and CRS, and a possible conflict between IIT and some versions of PIT—*e.g.*, Mendelovici and Bourget's (2020) variant of PIT—that reject holism and claim instead that contents depend on local properties of a state.

However, I believe that interpreting IIT as a form of CRS would be imprecise. This is because in CRS the nature of the content-fixing relations is *functional*, while in IIT the nature of the content-fixing relation is purely *structural* (Ellia et al. 2021): What fixes the contents is not what the system *does* in virtue of its relational profile, but simply that the relational profile *exists* in the first place. Another way to put this distinction is in terms of relevant content-fixing level: According to CRS, the relevant level is psychological, as it focuses on the role a *mental* state plays in a network, while, according to IIT, the relevant context-fixing level is physical,[11] as it focuses on the dispositional properties of the *neurons*. This means that, in IIT, once we have fixed the dispositional properties of a physical system, we have fixed its mental contents—further constructs related to the cognitive architecture of the system, and the inferential/functional profile it allows, are irrelevant.

We have thus arrived at the result that IIT can be interpreted as a theory of content: A version of PIT that I call 'structuralist PIT'. This is the idea that mental content is explanatorily parasitic on consciousness, being mental content identical to integrated information states (as cause-effect structures) and being integrated information states explanatorily identical to consciousness. This view is structuralist because conscious states are essentially relational, and correspond to the cause-effect structure fixed by the dispositional properties of a physical substrate. Again, IIT as structuralist PIT is non-representational, given that cause-effect structures are not representations of external properties, and therefore IIT's version of structuralist PIT should not be conflated with structuralist accounts of representational content (Shea 2018), nor with representation-based structuralist accounts to study consciousness (Lyre 2022). IIT's structuralist PIT is a strong version of PIT, given that all intentional states occur *qua* integrated information states, and by IIT's postulation integrated information states are phenomenal states. Moreover, IIT's structuralist PIT is firmly internalist, since the content-fixing conditions are defined by the boundaries of a Φ^{Max}-generating system.

This discussion has shown that IIT implies a version of PIT, as stated by P3. This means that an IIT-inspired version of PIT can be adopted to establish PIT through a neuroscientifically respected framework, and can be perhaps used to

[11] As stated above, in IIT, the "physical" is an operational construct posited from within consciousness itself. Evaluating the ontological implications of this positions is outside the scope of this paper, but see Tononi et al. 2022 for a discussion.

solve some of the most pressing problems for PIT. Mendelovici and Bourget (2020) survey four main challenges for PIT: Thoughts, wide intentional states (*i.e.*, states whose contents seem to be partly grounded on states of the external world), standing propositional attitudes (*i.e.*, my *belief that* Wellington is the capital of New Zealand, or my *desire that* Juventus win the Champions League), and non-conscious representational states. An interesting question, to which I will now (partly) turn, is how IIT could help with these challenges. In my argument, P4 claims that IIT can in fact do that, and I will now give an example of how that could be the case.

The first three problems individuated by Mendelovici and Bourget (2020) are mainly philosophical, while the fourth directly clashes with standard scientific practice. Since IIT is primarily a scientific theory of consciousness, in this paper I will limit my analysis to the specific question of how IIT as structuralist PIT can address that fourth challenge, that of accounting for non-conscious representational states often posited in scientific practice.

6. IIT and the Challenge of Non-Conscious Representational States

Non-conscious representational states seem to be at work in cases such as blindsight (Weiskrantz 1986), *i.e.*, when a patient reports seeing nothing on one half of the visual field despite being able to guess with accuracy well above chance what is present in that half of the visual field. Similar phenomena can be experimentally induced, courtesy of, for example, the subliminal priming effect stimulated by methodologies like visual masking. This happens when a target stimulus is rendered invisible by the presentation of another brief stimulus, presented immediately before and/or after the target (Kouider and Dehaene 2007). In general, the standard view in cognitive neuroscience seems to take the impressive amount of works showing the effects of subliminal perception on behaviour to support the view that there can be intentional (*i.e.*, representational, in this case) states that are not conscious.

To explore this point, let us consider the experiment by (Henson, Mouchlianitis, Matthews, and Kouider 2008), which uses a masking paradigm to study masked face priming. Part of this study is focused on the fame judgment subjects make after perceiving a face: Subjects have to report whether the face presented is a familiar or unfamiliar face. The study shows that the reaction time is significantly faster for "subliminally primed" faces, namely when the face is preceded by the same masked face,[12] with the priming effect being larger for familiar faces.

This result seems to support the idea that there can be intentional states that are not phenomenal states: The states probed by the mask seem to be *about a face*, because faster reaction times seem to be explained by the previous encounter with that face, but participants did not report being conscious of that face. This seems to falsify the basic tenet of strong PIT, to which, if the present analysis is on the right track, IIT subscribes.

[12] Participants were not aware of the mask, and were later tested on whether they consciously perceived the mask or not. The results show that subjects did not report consciously perceiving the face.

The main point of this section is this: According to IIT, the non-conscious perception *of a face* is a perception of a face only from the point of view of an extrinsic observer, and not from the intrinsic perspective of the conscious subject—the non-conscious state is not genuinely a state *about a face*. This is because face-neurons in the fusiform face gyrus (FFA) and occipital face gyrus (OFA), in the case of masked face priming as described by Henson et al. (2008), might not make and take a difference to and from the other neurons constituting the Φ^{Max}-generating system. Perhaps face-neurons can be causally connected to the Φ^{Max}-generating system, and therefore they can be inputs to the consciousness-generating system, without actually being a constitutive part of that system, because the causal connection can be severed without loss of integrated information. Or perhaps the state of OFA/FFA neurons can constitute a Φ-generating system partially overlapping with the main Φ^{Max}-generating system, and feeding directly into output systems. But if the Φ of this system is not *maximal*, it cannot be consciousness-constituting, because of the exclusion postulate.

If this is correct, the activity of FFA and OFA neurons can correspond to the content "face" only for an external observer that reconstructs the relation between neuronal activity and behaviour, but from the intrinsic perspective the face content simply does not exist. Thus, IIT seems to have the resources to claim that the non-conscious states involved in cases like masked face priming are simply *not* about a face, from the intrinsic perspective. The attribution of non-conscious content is instead totally dependent on an external point of view. This result can be generalized to any non-conscious state (either experimentally induced, or due to pathologies and psychiatric conditions) that the standard view would consider genuinely *intentional*. IIT's answer is that genuine content is intrinsic content, and therefore any purported non-conscious content, if not experienced from the first-person perspective, simply does not exist in any strong sense. Armed with this theoretical package, IIT can refute the idea that non-conscious representational states are a challenge to the main tenet of strong PIT, namely that all intentional states are phenomenal states, simply by rejecting that non-conscious representational states are genuine intentional states.[13]

If this analysis is correct, IIT seems to be able to aid PIT by providing a neuroscientifically-informed framework to respond to the challenge of non-conscious representational states, and thus by contributing to the strength of PIT as such.

7. Conclusion

In this paper, I have argued that IIT implies a version of phenomenal intentionality theory about content that I have labelled 'structuralist PIT'. This means that, since IIT is the only neuroscientific theory of consciousness that more or less explicitly subscribes to PIT, if IIT turns out to be more empirically adequate than other neuroscientific theories of consciousness, then PIT would have the upper hand against representationalism.

Here, I have not defended the claim that IIT is in fact the best available theory of consciousness, and perhaps the series of experiments based on adversarial collaboration currently ongoing will help shed light on this issue in the long term (Melloni et al. 2021).

[13] This strategy seems to be prima facie compatible with the eliminativist strategy in Mendelovici and Bourget 2020.

Further research can clarify the exact details of IIT as structuralist PIT, in order to cash out IIT as a more precise theory of content. This could help see how exactly IIT as structuralist PIT can deal with the other traditional challenges for PIT, from wide content to thoughts and standing propositional attitudes.

This paper has argued that a higher degree of confidence in IIT should be accompanied by a higher degree of confidence in PIT. Because of this, I argue that PIT theorists about content should see IIT favourably, because the empirical successes of IIT could reflect on the debate about content by supporting PIT.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A.M., Marshall, W., Mayner, W.G.P., Zaeemzadeh, A., Boly, M., Juel, B.E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J.P., Tononi, G. 2023, "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms", *PLoS Computational Biology* 19, 10, e1011465, DOI: 10.1371/journal.pcbi.1011465

Albantakis, L., Hintze, A., Koch, C., Adami, C., and Tononi, G. 2014, "Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity", *PLoS Computational Biology*, 10, 12, e1003966, DOI: 10.1371/journal.pcbi.1003966

Baker, B., Lansdell, B., and Kording, K.P. 2022, "Three aspects of representation in neuroscience", *Trends in Cognitive Sciences*, 26, 11, 942-958, DOI: doi.org/10.1016/j.tics.2022.08.014

Balduzzi, D. and Tononi, G. 2009, "Qualia: The Geometry of Integrated Information", *PLoS Computayional Biology*, 5, 8, e1000462, DOI: 10.1371/journal.pcbi.1000462

Barbosa, L.S., Marshall, W., Streipert, S., Albantakis, L., and Tononi, G. 2020, "A Measure for Intrinsic Information", *Scientific Reports*, 10, 1, 18803, DOI: 10.1038/s41598-020-75943-4

Bayne, T. 2007, "Conscious States and Conscious Creatures: Explanation in the Scientific Study of Consciousness", *Philosophical Perspectives*, 21, 1, 1-22, DOI: 10.1111/j.1520-8583.2007.00118.x

Block, N. 1995, "On a Confusion About a Function of Consciousness", *Behavioral and Brain Sciences*, 18, 2, 227-47.

Block, N. 1998, "Conceptual Role Semantics", in Craig, E. (ed.), *Routledge Encyclopedia of Philosophy*, New York: Routledge, 242-56.

Boly, M., Massimini, M., Tsuchiya, N., Postle, B.R., Koch, C., and Tononi, G. 2017, "Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence", *Journal of Neuroscience*, 37, 40, 9603-9613, DOI: 10.1523/JNEUROSCI.3218-16.2017

Bourget, D. and Mendelovici, A. 2016, "Phenomenal Intentionality", *The Stanford Encyclopedia of Philosophy*.

Brown, R., Lau, H., and LeDoux, J.E. 2019, "Understanding the Higher-Order Approach to Consciousness", *Trends in Cognitive Sciences*, 23, 9, 754-68, DOI: 10.1016/j.tics.2019.06.009

Cea, I. 2021, "Integrated Information Theory of Consciousness is a Functionalist Emergentism", *Synthese*, 199, 2199-2224, DOI: 10.1007/s11229-020-02878-8

Dalbey, B., and Saad, B. 2022, "Internal Constraints for Phenomenal Externalists: A Structure Matching Theory", *Synthese*, 200, 5, 1-29, DOI: 10.1007/s11229-022-03829-1

Edelman, G.M. 1987, "Neural Darwinism: The Theory of Neuronal Group Selection", New York: Basic Books.

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., Lang, P., Haun, J., Albantakis, L., Boly, M., and Tononi, G. 2021, "Consciousness and the Fallacy of Misplaced Objectivity", *Neuroscience of Consciousness*, 2021, 2, niab032, DOI :10.1093/nc/niab032

Facchin, M. 2021, "Structural Representations do not Meet the Job Description Challenge", *Synthese*, 199, 5479-5508, DOI: 10.1007/s11229-021-03032-8

Fink, S.B., Kob, L., and Lyre, H. 2021, "A Structural Constraint on Neural Correlates of Consciousness", *Philosophy and the Mind Sciences*, 2, DOI: 10.33735/phimisci.2021.79

Fodor, J.A. 1990, *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.

Francken, J.C., Beerendonk, L., Molenaar, D., Fahrenfort, J.J., Kiverstein, J.D., Seth, A.K., and van Gaal, S. 2022, "An Academic Survey on Theoretical Foundations, Common Assumptions and the Current State of Consciousness Science", *Neuroscience of Consciousness*, 2022, 1, niac011, DOI: 10.1093/nc/niac011

Gładziejewski, P. 2015, "Explaining Cognitive Phenomena with Internal Representations: A Mechanistic Perspective", *Studies in Logic, Grammar and Rhetoric*, 40, 1, 63-90, DOI: 10.1515/slgr-2015-0004

Grasso, M. 2019, "IIT vs. Russellian Monism: A Metaphysical Showdown on the Content of Experience", *Journal of Consciousness Studies*, 26, 1-2, 48-75.

Grasso, M., Haun, A.M., and Tononi, G. 2021, "Of Maps and Grids", *Neuroscience of Consciousness*, 2021, 2, niab022, DOI: 10.1093/nc/niab022

Haun, A.M. and Tononi, G. 2019, "Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience", *Entropy*, 21, 12, 1160, DOI: 10.3390/e21121160

Henson, R.N., Mouchlianitis, E., Matthews, W.J., and Kouider, S. 2008, "Electrophysiological Correlates of Masked Face Priming", *Neuroimage*, 40, 2, 884-95, DOI: 10.1016/j.neuroimage.2007.12.003

Horgan, T. and Tienson, J. 2002, "The Phenomenology of Intentionality and the Intentionality of Phenomenology", in Chalmers, D.J. (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford: Oxford University Press, 520-33.

Hutto, D. and Myin, E. 2013, *Radicalizing Enactivism: Basic Minds Without Content*, Cambridge, MA: MIT Press.

Kim, J. 1998, *Current Issues in Philosophy of Mind*, Cambridge: Cambridge University Press.

Kouider, S. and Dehaene, S. 2007, "Levels of Processing During Non-Conscious Perception: A Critical Review of Visual Masking", *Philosophical Transactions of the Royal Society B Biological Sciences*, 362, 1481, 857-75, DOI: 10.1098/rstb.2007.2093

Kriegel, U. 2013, *Phenomenal Intentionality*, Oxford: Oxford 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

Lee, J. 2019, "Structural Representation and the Two Problems of Content", *Mind and Language*, 34, 5, 606-26.

Lyre, H. 2022, "Neurophenomenal Structuralism: A Philosophical Agenda for a Structuralist Neuroscience of Consciousness", *Neuroscience of Consciousness*, 1, niac012, DOI: 10.1093/nc/niac012

Márton, M. 2022, "Intentional and Phenomenal Properties: How Not to Be Inseparatists", *Philosophia*, 50, 1, 127-47, DOI: 10.1007/s11406-021-00362-2

Mashour, G.A., Roelfsema, P., Changeux, J.-P., and Dehaene, S. 2020, "Conscious Processing and the Global Neuronal Workspace Hypothesis", *Neuron*, 105, 5, 776-98, DOI: 10.1016/j.neuron.2020.01.026

McKilliam, A.K. 2020, "What is a Global State of Consciousness?", *Philosophy and the Mind Sciences*, 1, II, DOI: 10.33735/phimisci.2020.II.58

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D.J., Devore, S., Fallon, F., de Lange, F.P., Jensen, O., Kreiman, G., Luo, H., Panagiotaropoulos, T.I., Dehaene, S., Koch, C., and Tononi, G. 2023, "An Adversarial Collaboration Protocol for Testing Contrasting Predictions of Global Neuronal Workspace and Integrated Information Theory", *PLoS ONE*, 18, 2, e0268577, DOI: 10.1371/journal.pone.0268577

Melloni, L., Mudrik, L., Pitts, M., and Koch, C. 2021, "Making the Hard Problem of Consciousness Easier", *Science*, 372, 6545, 911-12, DOI: 10.1126/science.abj3259

Mendelovici, A. 2018, *The Phenomenal Basis of Intentionality*, Oxford: Oxford University Press.

Mendelovici, A. and Bourget, D. 2020, "Consciousness and Intentionality", in Kriegel, U. (ed.), *The Oxford Handbook of the Philosophy of Consciousness*, Oxford: Oxford University Press, 560-85.

Millikan, R. 1989, "Biosemantics", *The Journal of Philosophy*, 86, 6, 281-97.

Mindt, G. 2021, "Not All Structure and Dynamics Are Equal", *Entropy*, 23, 9, 1226, DOI: 10.3390/e23091226

Nagel, T. 1974, "What Is It Like to Be a Bat?", *The Philosophical Review*, 83, 4, 435-50, DOI: 10.2307/2183914

Neander, K. 2008, "Teleological Theories of Mental Content: Can Darwin Solve the Problem of Intentionality?", in Ruse, M. (ed.), *The Oxford Handbook of Philosophy of Biology*, Oxford: Oxford University Press, 381-409.

Neander, K. 2017, *A Mark of the Mental: A Defence of Informational Teleosemantics*, Cambridge, MA: MIT Press.

Negro, N. 2022, "Emergentist Integrated Information Theory", *Erkenntnis*, 1-23. DOI: 10.1007/s10670-022-00612-z

O'Regan, K. 2014, "The Explanatory Status of the Sensorimotor Approach to Phenomenal Consciousness, and Its Appeal to Cognition", in Bishop, J.M., and Martin, A. (eds.), *Contemporary Sensorimotor Theory*, *Studies in Applied Philosophy, Epistemology and Rational Ethics* 23, Cham: Springer, 23-35.

Oizumi, M., Albantakis, L., and Tononi, G. 2014, "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0", *PLoS Computational Biology*, 10, 5, e1003588, DOI: 10.1371/journal.pcbi.1003588

Pautz, A. 2020, "Representationalism about Consciousness", in Kriegel, U. (ed.), *The Oxford Handbook of the Philosophy of Consciousness*, Oxford: Oxford University Press, 405-37.

Ramsey, W.M. 2007, *Representation Reconsidered*, Cambridge: Cambridge University Press.

Schlicht, T. 2018, “Does Separating Intentionality from Mental Representation Imply Radical Enactivism?”, *Frontiers in Psychology*, 9, DOI: 10.3389/fpsyg.2018.01497

Searle, J.R. 1980, “Minds, Brains, and Programs”, *Behavioral and Brain Sciences*, 3, 3, 417-24, DOI: 10.1017/S0140525X00005756

Searle, J.R. 1992, *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

Searle, J.R. 2013, “Can Information Theory Explain Consciousness”, *The New York Review of Books*, https://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/

Seth, A.K. and Baars, B.J. 2005, “Neural Darwinism and Consciousness”, *Consciousness and Cognition*, 14, 1, 140-68, DOI: 10.1016/j.concog.2004.08.008

Seth, A.K. and Bayne, T. 2022, “Theories of Consciousness”, *Nature Reviews Neuroscience*, DOI: 10.1038/s41583-022-00587-4

Shea, N. 2018, *Representation in Cognitive Science*, Oxford: Oxford University Press.

Signorelli, C.M., Szczotka, J., and Prentner, R. 2021, “Explanatory Profiles of Models of Consciousness—Towards a Systematic Classification”, *Neuroscience of Consciousness*, 2021, 2, niab021, DOI: 10.1093/nc/niab021

Tononi, G. 2004, “An Information Integration Theory of Consciousness”, *BMC Neuroscience*, 5, 42, DOI: 10.1186/1471-2202-5-42

Tononi, G. 2008, “Consciousness as Integrated Information: A Provisional Manifesto”, *Biological Bulletin*, 215, 3, 216-42, DOI: 10.2307/25470707

Tononi, G. 2012, “Integrated Information Theory of Consciousness: An Updated Account”, *Archives Italiennes de Biologie*, 150, 4, 293-329.

Tononi, G. 2015, “Integrated Information Theory”, *Scholarpedia*, 10, 1, 4164, DOI: 10.4249/scholarpedia.4164

Tononi, G., Albantakis, L., Boly, M., Cirelli, C., and Koch, C. 2022, “Only What Exists Can Cause: An Intrinsic View of Free Will”, *arXiv*, DOI: 10.48550/ARXIV.2206.02069

Tononi, G., Boly, M., Massimini, M., and Koch, C. 2016, “Integrated Information Theory: From Consciousness to its Physical Substrate”, *Nature Reviews Neuroscience*, 17, 7, 450-61, DOI: 10.1038/nrn.2016.44

Tononi, G. and Koch, C. 2015, “Consciousness: Here, There and Everywhere?”, *Philosophical Transactions of the Royal Society B Biological Sciences*, 370, 1668, DOI: 10.1098/rstb.2014.0167

Tononi, G., Sporns, O., and Edelman, G.M. 1996, “A Complexity Measure for Selective Matching of Signals by the Brain”, *Proceedings of the National Academy of Sciences of the United States of America*, 93, 8, 3422-27, DOI: 10.1073/pnas.93.8.3422

Tsuchiya, N. 2017, ““What Is It Like to Be a Bat?”—A Pathway to the Answer from the Integrated Information Theory”, *Philosophy Compass*, 12, 3, e12407, DOI: 10.1111/phc3.12407

Tsuchiya, N., and Saigo, H. 2021, “A Relational Approach to Consciousness: Categories of Level and Contents of Consciousness”, *Neuroscience of Consciousness*, 2021, 2, niab034, DOI: 10.1093/nc/niab034

Tsuchiya, N., Taguchi, S., and Saigo, H. 2016, “Using Category Theory to Assess the Relationship between Consciousness and Integrated Information Theory”, *Neuroscience Research*, 107, 1-7, DOI: 10.1016/j.neures.2015.12.007

Tye, M. 1995, *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*, Cambridge, MA: MIT Press.

Weiskrantz, L. 1986, *Blindsight*: *A Case Study and Implications*, Oxford: Oxford University Press.

Yaron, I., Melloni, L., Pitts, M., and Mudrik, L. 2022, "The ConTraSt Database for Analysing and Comparing Empirical Studies of Consciousness Theories", *Nature Human Behaviour*, 6, 4, 593-604, DOI: 10.1038/s41562-021-01284-5

Zeleznikow-Johnston, A., Aizawa, Y., Yamada, M., Tsuchiya, N. 2023, "Are Color Experiences the Same across the Visual Field?", *Journal of Cognitive Neuroscience,* 35, 4, 509-42, DOI: https://doi.org/10.1162/jocn_a_01962