The Integrated Information Theory of consciousness: A case of mistaken identity

Authors:

Bjorn Merker
Independent scholar. Fjälkestadsv. 410-82, 29194 Kristianstad, Sweden
bjornmerker@gmail.com

Kenneth Williford
Department of Philosophy & Humanities, University of Texas at Arlington
williford@uta.edu https://mentis.uta.edu/explore/profile/kenneth-williford

David Rudrauf
Faculty of Psychology and Education Science, Swiss Center for Affective Science, University Center of Computer Science, University of Geneva, Geneva, Switzerland
David.Rudrauf@unige.ch https://www.unige.ch/fapse/mmef/en/recherche/

Corresponding author: Bjorn Merker https://orcid.org/0000-0003-2941-6133

Word counts:
Short abstract: 94
Short Abstract

Giulio Tononi’s Integrated Information Theory (IIT) is a prominent bid for a scientific explanation of consciousness based on the direct identification of consciousness with integrated information. This has the consequence of requiring the attribution of consciousness to electrical power grids, gene expression networks, and social networks, *inter alia*. IIT is thus led towards the orbit of panpsychist ideation. We examine IIT’s identity claim in its formal, phenomenological, and neuroscientific aspects, concluding that it fails as a theory of consciousness, a failure that harbors lessons for continued progress towards a scientific account of the phenomenon. [94 words]

Long Abstract

Giulio Tononi’s Integrated Information Theory (IIT) proposes explaining consciousness by directly identifying it with integrated information. We examine the construct validity of IIT’s measure of consciousness, *phi* (Φ), by analyzing its formal properties, its relation to key aspects of consciousness, and its co-variation with relevant empirical circumstances. Our analysis shows that IIT’s identification of consciousness with the causal efficacy with which differentiated
networks accomplish global information transfer (which is what \( \Phi \) in fact measures) is mistaken. This misidentification has the consequence of requiring the attribution of consciousness to a range of natural systems and artifacts that include, but are not limited to, large-scale electrical power grids, gene-regulation networks, some electronic circuit boards, and social networks. Instead of treating this consequence of the theory as a disconfirmation, IIT embraces it. By regarding these systems as bearers of consciousness \textit{ex hypothesi}, IIT is led towards the orbit of panpsychist ideation. This departure from science as we know it can be avoided by recognizing the functional misattribution at the heart of IIT’s identity claim. We show, for example, what function is actually performed, at least in the human case, by the cortical combination of differentiation with integration that IIT identifies with consciousness. Finally, we examine what lessons may be drawn from IIT’s failure to provide a credible account of consciousness for progress in the very active field of research concerned with exploring the phenomenon from formal and neural points of view.

1. Introduction

We may somewhat arbitrarily date the beginning of contemporary neuroscientific interest in consciousness to 1949 and the discovery of the cerebral activating and alerting functions of the brainstem’s reticular formation (Moruzzi & Magoun 1949). This discovery sparked a decade of pioneering research and theory summarized in major symposium volumes that featured the term ‘consciousness’ in their titles: \textit{Brain mechanisms and consciousness} (Adrian et al., eds.)
1954) and *Brain and conscious experience* (Eccles, ed. 1966). Though the hopes pinned on the reticular activating system as the key to consciousness went unfulfilled, in subsequent decades the topic of consciousness found exponents in psychology (Shallice 1972; Luria 1973; Mandler 1975; Baars 1988) as well as in behavioral biology (Griffin 1976; Schwartz & Shapiro, eds. 1978) and elsewhere. Since then, a wide variety of novel bids for a neurobiological account of consciousness have appeared. They include proposals regarding re-entrant processing (Edelman 1989), neuronal correlates of subjective experience (Logothetis & Schall 1989), gamma synchrony (Crick & Koch 1990; Llinas et al. 1998; Singer 1998); the “neural self” and core consciousness (Damasio 1998); a neuronal global workspace (Dehaene et al. 1998), neural complexity (Tononi & Edelman 1998), feedforward-feedback convergence (Lamme 2000), a neural reality space (Merker 2007); integrated information (Tononi 2008), selfhood and interoception (Seth et al. 2011), the entropic brain (Carhart-Harris et al. 2014), and “projective consciousness” (Rudrauf et al. 2017), among many others.

Contemporary science is, in other words, well embarked on a quest to obtain closure at the observer end of its naturalistic project by finding a neurobiological or (neuro)computational account of the conscious observer itself. The heterogeneity of the just-cited sample of candidate accounts indicates, however, that the quest is still in its early stages. In ways that frustratingly echo the results of past attempts to approach consciousness seriously from the 17th century to the 19th, and this in spite of the prodigious accumulation of relevant knowledge in the 20th and 21st, the quest has yet to crystallize around a generally agreed-upon paradigm featuring shared first principles, *explananda*, evidentiary criteria, and methodology. As a result, “schools”
advocating disparate theoretical bids for paradigmatic status contend with one another without the benefit of any shared means of adjudicating the relative merits of their conflicting proposals. In Kuhnian terms, these are the cardinal characteristics of the “prehistory” stage of a nascent science (Kuhn 1970, Sect. 2). On this “route to normal science,” the proponents of rival theories, formulated in scientific terms and supported by empirical evidence, often find themselves talking past one another for lack of shared assumptions. Such a scientifically unsatisfactory state of affairs is typically resolved, still speaking with Thomas Kuhn, at the transition to science proper “…by the triumph of one of the pre-paradigm schools, which, because of its own characteristic beliefs and preconceptions, emphasized only some special part of the too sizable and inchoate pool of information” (Kuhn 1970, p. 17).

One of the research programs currently aspiring to this paradigmatic position is the body of work constituting the Integrated Information Theory (IIT) of consciousness, propounded by Giulio Tononi and colleagues (Tononi 2004, 2008, 2012; Oizumi et al. 2014; Tononi et al. 2016; Mayner et al. 2018). As a candidate paradigm for a scientific account of consciousness, it deserves thorough scrutiny in its conceptual, formal, and empirical aspects. This is all the more so because, according to IIT, consciousness is not uniquely associated with animal brains and their potential creations (e.g., artificial consciousness in machines) but extends, as we shall see, to natural objects and artifacts much farther afield. As a consequence, aspects of panpsychist metaphysics have begun to take on the appearance of relevance to the conceptual framework of IIT (Tononi and Koch 2015), a development seized upon by the popular science media under headlines such as “Minds everywhere: ‘panpsychism’ takes hold in science” (Ghose 2016).
In what follows, we motivate our claim that IIT rests upon a misconstrual of what its central concept of integrated information and prime quantity $\Phi$ in fact reflect. First, we examine the construct validity of IIT’s measure of consciousness, $\phi$ ($\Phi$), by placing it in formal, conceptual, and empirical contexts in Sections 2, 3 and 4. They address in turn the construct ‘integrated information’ in terms of its formal properties, its relation to salient, possibly defining aspects of consciousness, and its co-variation with relevant empirical circumstances. In Section 5, we examine how its identification of what $\Phi$ measures with consciousness has brought IIT towards the orbit of panpsychist ideation. Section 6, finally, explores lessons that may be gleaned from the manner in which IIT fails as a theory of consciousness. These lessons concern the constitution of consciousness and also bear on the current impasse in the search for its neural correlates. We end by suggesting that better articulated and constrained hypotheses about the nature and structure of consciousness are required to enable continued progress towards an empirically satisfactory theory and offer a few suggestions in this regard.

2. IIT in context, I: integrated information

IIT identifies consciousness with integrated information, that is, according to IIT, consciousness just is integrated information, a constitutive claim: “According to the IIT, consciousness is one
and the same thing as integrated information” (Tononi 2008, p. 232). Much therefore hinges on how integrated information, and its measure, $\Phi$, are understood in IIT. In its latest version, integrated information is construed as the extent to which the causal structures of a system, e.g. the brain, achieves state transitions that cannot be reduced to the joint transitions of their parts independently of one another (Oizumi et al. 2014). More colloquially expressed, integrated information is “...the amount of information generated by a complex of elements, above and beyond the information generated by its parts” (Tononi 2008, p. 216). Such an informational difference is an attribute of those parts of a system whose causal structure exhibits a kind of complexity that combines system differentiation with integration (Tononi, Sporns and Edelman 1994; Tononi and Edelman 1998, 2000, Ch. 6; Mediano et al. 2019b). That is the bedrock on which IIT rests its identification of integrated information with consciousness (Tononi 2004, 2008, 2012; Oizumi et al. 2014).

IIT not only identifies consciousness with integrated information, it proposes a formal (algorithmic) approach for quantifying the degree to which a physical system exhibits integrated information and thus, ex hypothesi, consciousness (Mayner et al. 2018). A system here is understood as a set of potentially active nodes and their interconnections, i.e., a network in the graph-theoretic sense. Without entering too heavily into technical details, the key operation in the way IIT measures the extent to which such a system combines integration with differentiation is first to analyze the behavior, under iterated perturbation, of each of its subsets

---

1 Further statements of this identity are: “…an experience is a maximally integrated conceptual information structure”, and “…a local maximum of integrated information is indeed identical with consciousness, as claimed by IIT” (Tononi, 2012, p. 298, 309); “…according to IIT, there is an identity between phenomenological properties of experience and informational/causal properties of physical systems” measured by $\Phi$, and “…only mechanisms that specify integrated information can contribute to consciousness” (Oizumi et al., 2014, pp. 3, 8).
(S₁...Sn) of nodes (or elements) in terms of the quantification of information transfer across all the bipartitions of the subset (Si) under consideration (a bipartition divides Si into one of its subsets and its complement relative to Si), in each direction and across each bipartition. The extent to which information transfer within such a subset (Si) is not accounted for by the separate behaviors of its bipartitions (its “irreducibility” to their joint behavior, in IIT terms) is the extent to which that subset (Si) integrates information, as measured by a scalar quantity Φ (see, e.g., Steinhart 2018, pp. 160-162 for an accessible technical presentation). There are many circumstances in which such a procedure would identify multiple system subsets with substantial Φ. On IIT’s identification of consciousness with integrated information this would entail the coexistence of multiple consciousnesses in a single system at a given time, rather than a single, unified consciousness, as phenomenology and parsimony considerations would seem to dictate. IIT eliminates the possibility of multiple consciousnesses in a single system by fiat through an “Axiom of Exclusion”² that stipulates that only the subset with the largest Φ, or ΦMAX (the “maximally irreducible” one) is the conscious one.

Additional formalities concern the manner in which nodes are perturbed, the way to deal with the distorting consequences of very unequal bipartitions, methods for assessing functional interdependence (information transfer) across a bipartition, and whether one or two time steps are included in the calculations. Though choices in these respects have varied in different versions of IIT, they need not detain us here because they are not germane to the arguments.

² In the latest version of IIT this axiom is formulated as follows: “Consciousness is exclusive: each experience excludes all others – at any given time there is only one experience having its full content, rather than a superposition of multiple partial experiences; each experience has definite borders – certain things can be experienced and others cannot; each experience has a particular spatial and temporal grain – it flows at a particular speed, and it has a certain resolution such that some distinctions are possible and finer or coarser distinctions are not” (Oizumi et al. 2014, p. 3).
that follow. We are concerned, rather, with the generic core of the IIT formalism, which is
designed to capture the extent to which a network combines differentiation with integration in
the above sense. That core has remained invariant irrespective of formal differences between
IIT’s versions, as has the claim that integrated information (described and assessed in a
particular way) is consciousness. Our concern with the generic core of IIT is twofold: IIT relies on
an under-constrained characterization of consciousness, a matter to be considered in Sections 3
and 6; moreover, integrated information itself is a hypothetical construct in search of a unique
and well-defined formal definition and measure (Tegmark 2016; Barrett & Mediano 2019;
Mediano et al. 2019a; Mediano et al. 2019b). As is also the case for the related construct
“dynamical complexity” (see Mediano et al., 2019a, b), “integrated information” remains a
“work in progress”.

The current version of IIT is formally defined only for discrete and memoryless (Markovian)
systems (Balduzzi & Tononi 2008; Oizumi et al. 2014). But even under these simplifying
assumptions, the computational requirements for determining $\Phi$ grow exponentially with the
number of elements in the system, preventing it from being implemented and analyzed on
anything more than toy systems of less than 20 elements, and typically far fewer.\(^3\)

Compounding the computational burden on IIT is the fact that it must search for a maximal $\Phi$
($\Phi^{\text{MAX}}$) over all scales at which a system’s “elements” might be defined (Tononi 2004; Oizumi et
al. 2014; Tegmark 2016). In effect, IIT, by its own criteria, identifies the conscious subset of

\(^3\) This is to be compared with the number of “elements” composing the human brain. For the cerebral cortex
alone this number ranges anywhere from its total of 360 areas (Glasser et al. 2016) – if whole cortical areas are
treated as “elements” – to millions of cortical functional columns, and twenty billion cortical neurons (Pakkenberg
and Gundersen 1997), not to speak of the finer scales over which IIT in principle requires a search to be conducted.
elements *a posteriori*, after $\Phi^{\text{MAX}}$ is found. For the brain, this means that, in principle, the search must be conducted from at least the level of its atoms to that of whole cortical areas (Balduzzi and Tononi 2008; Tegmark 2016). This compounded computational intractability alone prevents IIT from being tested on relevant brain data.\(^4\) While issues of computational intractability may indeed limit the practical applicability of parameter-adequate models to certain complex physical systems, it is hardly common to leave matters so wide open, vis-à-vis relevant “elements”, as to span from (at least) atoms to large-scale populations of neurons.

As a result of these various obstacles to empirical testing, a number of proxy measures for $\Phi$ have been developed (Barrett and Seth 2011; Mediano et al. 2019b). However, relating these to $\Phi$ is problematic because even variant “…measures that share similar theoretical properties can behave in substantially different ways, even on simple systems” (Mediano et al. 2019b, p. 23). For example, variant measures of integrated information do not agree on the relative ranking of integration scores across the members of a diverse set of network types (Mediano et al. 2019b, Figs. 4 and 5). This seems to undermine the validity of current proxy measures as stand-ins for $\Phi$ and counsels extreme caution in attempting to relate experimental results obtained by proxy measures to the theoretical claims of IIT.

In this discouraging situation, much would be gained if there were a way of exploring integrated information apart from the formalism of IIT. A way to do so has in fact been provided by the recent development of a method that removes the principal computational bottleneck in

\(^4\) A further complication in this regard is that physiological data typically come in time-series of continuous variables, while IIT, as currently formulated, assumes discreteness.
measuring integrated information, viz., the problem of finding the so-called “minimum information bipartition” within a system (Toker & Sommer 2019). This construct, also known as “the cruelest cut”, is the terminus of the search across bipartitions for the one with the least causal interdependence across the partition. It is this crucial step in determining $\Phi$ that is subject to a combinatorial explosion over the number of elements of the system. Toker and Sommer have remedied this computational bottleneck by applying the graph-theoretic method of “spectral clustering” to the search problem (Toker and Sommer 2019; for a concise graph theory primer, see Stam n.d.). Integrated information can now be assessed on large arrays of neural time series data, as demonstrated by a proof-of-principle analysis of electrocorticogram time series data from 125 electrodes distributed across one cerebral hemisphere in each of two macaques (Toker and Sommer 2019).

The Toker-Sommer advance sheds much needed light on the IIT formalism by showing that its measure of integrated information, $\Phi$, tracks the extent to which differentiated dynamic networks achieve efficiency in their global information transfer (Toker & Sommer 2019, Fig. 5, p. 16). For ease of reference we shall refer to such networks generically as efficient networks, irrespective of the details of how they combine integration with differentiation, and with the proviso that ‘efficiency’ here refers strictly to the efficiency with which a differentiated network accomplishes global information transfer. $\Phi$, it turns out, is one measure of this network property, a property that has been under intense scrutiny for decades in the discipline of network science. That discipline had its first beginnings in the mathematical analysis of random

---

5 IIT uses bipartition for this purpose, though, as pointed out by Scott Aaronson, there is no theoretical rationale for this choice (Aaronson 2014). The choice is pragmatic, however, since any other option aggravates IIT’s combinatorial problem.
graphs on the one hand (Erdős & Rényi 1959; Gilbert, 1959), and in the “small world”
connectivity of social networks on the other (Milgram 1967). It came into its own with
pioneering studies of network dynamics (Watts & Strogatz 1998; Barzel & Barabási 2013), the
scaling behavior of networks (Barabási & Albert 1999), and the development of a measure of
the efficiency of their global information transfer (Latora & Marchiori 2001).

Since then, network science has found a profusion of applications in the analysis of complex
systems. These include the logistics of transportation, telecommunication systems, and
computer networks, biological networks such as metabolic, immune and gene regulation
networks; neural, cognitive and semantic networks, and social networks (all reviewed by
Boccaletti et al. 2006; for neuroscience applications, see Sporns & Zwi 2004; Bullmore & Sporns,
et al. 2013). Further significant applications extend to evolutionary and ecological dynamics
(Lieberman et al. 2005), economics (Vitali et al. 2011), and epidemiology (Zhang & Zhen 2012).
In all these cases, the efficient networks, variants of which go under names such as “small
world”, “expander graph”, “rich club”, “core-periphery” and “bow-tie” networks, combine
integration with differentiation, the key construct of IIT. They do this as a means of performing
complex network functions under temporal, spatial, and energy efficiency constraints;
consciousness is not, to all appearances, a necessary condition of such network organization.

At the outset of his path to IIT, Tononi himself did not identify the extent to which a system
combines differentiation and integration with consciousness specifically. He had developed a
quantitative measure of this network attribute in work on the functional organization of the cerebral cortex with Olaf Sporns and Gerald Edelman (Tononi, Sporns & Edelman 1994). They called this immediate predecessor of $\Phi$ “neural complexity”, envisaging its broad relevance to biological and technological systems far afield from brains:

A number of other biological systems exhibit complexity and appear to be susceptible to the kind of analysis described here. The circuits of gene regulation in prokaryotes and eukaryotes, various endocrine loops, and the coordinative events observed during embryological development are significant examples. It remains to be seen whether our approach will also prove useful in more widespread applications such as the analysis of parallel computation and communication networks. (Tononi, Sporns & Edelman 1994, p. 5037).

They correctly anticipated, in other words, the relevance of their complexity measure to the diverse systems utilizing information integration explored independently of IIT by network science in subsequent years. IIT was born when Tononi took the additional step of directly identifying integrated information with consciousness (Tononi 2004).

This direct identification of integrated information as measured by $\Phi$ with consciousness is the central constitutive claim of IIT. It goes well beyond, in other words, the more modest claim that consciousness as we know it requires network efficiency of the kind measured by $\Phi$ to support its rich contents. As we shall discuss in more detail in Section 5, this radical claim commits IIT to the conscious status of all systems with substantial $\Phi$ scores – from gene expression networks to large-scale electrical power grids – a conceptual burden that has, rather incongruously, tended to bring IIT towards the orbit of panpsychist ideation (Tononi 2008, p. 236; Tononi &

---

6 As late as 2003, Tononi & Sporns listed “…gene regulatory networks, ecological and social networks, computer architectures, and communication networks” as examples of systems whose typical combinations of integration with differentiation lend themselves to analysis by their measure of integrated information, now called $\Phi$ (Tononi & Sporns 2003).
Koch, 2015). Much is therefore at stake in assessing the warrant for this identity claim. This mandates close scrutiny of the arguments and evidence marshaled in its support. We accordingly turn next to an examination of the grounds on which IIT bases its identification of consciousness with integrated information.

3. IIT in context, II: consciousness

IIT bases its identification of consciousness with integrated information on the way in which key principles of the domains appear to coincide. As we saw in Section 2, integrated information requires a causal system to combine integration and differentiation. Throughout the development of IIT, consciousness is described as being both integrated (the venerable synchronic “unity” of consciousness in any given instance) and differentiated (the rich diversity of its contents or, roughly, the informativeness of its separate instances) (Tononi and Edelman 1998; Edelman and Tononi 2000, Chs. 3, 6, 10 & 11; Tononi, 2004, 2008, 2012). These descriptors, and even their combination, are perfectly legitimate, yet partial, characterizations of consciousness. However, their fit with our intuitions about consciousness by no means entails that they must be understood in the precise sense in which ‘integrated’ and ‘differentiated’ are defined in the IIT formalism.

The issue is a crucial one: first, because the basic identity claim of IIT hinges on it, and second, because recent versions of IIT claim that the entire formal apparatus by which IIT defines and
measures $\Phi$ is derived from the phenomenologically available attributes of consciousness via their axiomatization and conversion into postulates governing the implementation in systems of what is measured by the IIT algorithms (Tononi 2012; Oizumi et al., 2014). The viability of IIT therefore rests on the adequacy of its characterization of consciousness and more specifically on whether the kind of “integration” and “differentiation” that can be ascribed to consciousness on phenomenological grounds, on the basis of first-person experience, is in fact the kind of integration and differentiation formalized in IIT. Unless, at the appropriate level of abstraction, the phenomenological characteristics are captured in all essentials by the formal constructs that define $\Phi$, there is no basis for claiming an identity between the prima facie disparate domains. In other words, the phenomenological predicates ‘integrated’ and ‘differentiated’ must pick out the very properties picked out by the homonymous predicates as defined in IIT. We shall examine IIT’s claims regarding the two attributes separately, beginning with ‘integration’, or “the unity of consciousness”.

The various arguments, illustrations, and thought experiments by which IIT seeks to identify consciousness with integrated information are introduced in compact form in the 1998 paper by Tononi and Edelman (and elaborated in more detail in Edelman and Tononi, 2000 and subsequent publications by Tononi). For “integration” as an attribute of consciousness they cite William James on its unity, followed by the claim that “Integration is best appreciated by

---

This “axiomatization” of IIT is based, first and foremost, on the two attributes picked out by the terms ‘integrated’ and ‘differentiated’, as we learn from an endnote in a paper written prior to the “axiomatic” development: “One could say that the theory starts from two basic phenomenological postulates—(i) experience is informative; (ii) experience is integrated—which are assumed to be immediately evident [...]. In principle, the theory, including the mathematical formulation and its corollaries, should be derivable from these postulates” (Tononi 2008, p. 240, endnote 1). By 2012 IIT was presented in terms of axioms and linked postulates, though the axioms now numbered five (Tononi 2012).
considering the impossibility of conceiving of a conscious scene that is not integrated, that is, one which is not experienced \textit{from a single point of view}” (Tononi and Edelman 1998, p. 1846, emphasis added; cf. Edelman and Tononi 2000, pp. 23-29). Similarly, in papers on IIT proper: “Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and that is experienced \textit{from a single point of view}” (Tononi 2008, p. 219, emphasis added; and identically in Tononi, 2012, p. 295). That is, as late as 2012, IIT links the phenomenology of integration to a feature of consciousness that is not uncommonly considered fundamental and essential to its unity, namely its property of realizing a single point of view (cf. Nagel 1974; Merker 2012, 2013b; Rudrauf et al. 2017; see also Section 6). How and where then, in the formalisms of IIT, is this viewpoint underwriting the unity of consciousness implemented?

According to IIT: “Since information can only be integrated within a complex and not outside its boundaries, consciousness as information integration is necessarily subjective, private, and related to a single point of view or perspective” (Tononi 2004, p. 6; ‘a complex’ here refers to a subset of system elements that qualifies as conscious by IIT criteria). Similarly: “A complex, then, can be properly considered to form a single entity having its own, intrinsic ‘point of view’ (as opposed to being treated as a single entity from an outside, extrinsic point of view)” (Tononi 2008, p. 221). The intent of these claims is to tie three separate, albeit related, aspects of consciousness – privacy, subjectivity, and the realization of a point of view – to integrated information as implemented in IIT. However, these claims are problematic.
To show that a set of elements coheres informationally in the way specified by the IIT algorithm, and therefore that it yields a substantial Φ score, is only to show that the set causally participates in the operation of an efficient network (whatever function that network may serve). It is not to show that the set “exists as an entity from its own intrinsic perspective” (Oizumi et al., 2014, p. 4). There is no rationale for the latter claim inherent in the IIT formalism. Specifically, where in the formalism is perspective, let alone “its own perspective”, defined? These locutions from phenomenologically based discourse on consciousness that are introduced as attributes of a network constellation without warrant in the formalism itself. This attribution of an “intrinsic perspective” to a set of elements of an efficient network occurs 22 times in Tononi (2012), and 25 times in Oizumi et al. (2014). It gradually eclipses reference to the unity of consciousness under a “point of view” in the exposition of IIT, a locution that has disappeared altogether in the latest version of IIT (Oizumi et al., 2014). However, nothing substantively corresponding to the first-person phenomenological fact of the unity of consciousness under a point of view is to be found anywhere in the formalisms of IIT (cf. Section 6), a major point of mismatch between phenomenology and the IIT formalism. “Unity” in the sense of putting many things together in some (functionally) organized fashion, which occurs in many systems, both biological and artificial, simply does not seem to be the same thing as the unity under a point of view characteristic of consciousness (cf. Section 6).

Thus, IIT has dispensed with reference to the phenomenological unity of consciousness under a point of view, in any robust sense. And is has taken what remained of that idea and assimilated it to its ad hoc attribution of an “intrinsic perspective” to any system with a substantial Φ score.
In this regard, IIT, by fiat, trivializes the concept of perspective or point of view by reducing it to the notion of individuation in terms of Φ (cf. Cerullo 2018, p. 8). The issue is a crucial one for consciousness theory and will receive further consideration in Sections 5 and 6.

But IIT also rests its formalism-phenomenology identity claim vis-à-vis the unity of consciousness on a presumed experiential irreducibility principle, as follows: “Consciousness is integrated: each experience is (strongly) irreducible to non-interdependent components” (Oizumi et al. 2014, p. 3, “Integration Axiom”). While we would not think of denying the phenomenological reality of the Gestalt-type phenomena that partly motivate this axiom (see, esp., Edelman and Tononi 2000, p. 25), it is far from clear that one can safely assume that the kind of “unity” that corresponds to a substantial Φ score is ipso facto of this Gestalt type, or that the latter is a special case of the former. In fact, as Merleau-Ponty pointed out in 1945, the basic figure-ground structure isolated and characterized by the Gestalt psychologists presupposes a viewpoint or “point-horizon” structure (see Merleau-Ponty 2012, e.g., pp. 102-105; cf. Lehar 2003, pp. 68ff.); this structure, as noted, is not recoverable from the IIT formalism by itself.

We would further agree that consciousness typically does seem to involve multimodal integration, temporal integration, feature binding, the coherent combination of sensory and cognitive factors, and possibly other forms of “unity”, in addition to unity under a point of view (cf. Section 6). But our contention is that it is far from clear that the IIT formalism expresses the single core principle behind these possibly quite different aspects of unity or that their specific
structures are so readily accessible in first-person experience as to be reduced to the form of a single axiom purporting to be purely phenomenological in substance.

The wording of the (putatively purely phenomenological) Axiom of Integration suggests that it is meant to echo the specific characterization of irreducibility inherent in the IIT formalism.

Arguably, it parallels the way the “Axiom of Exclusion” eliminates subsets of $\Phi^{\text{MAX}}$ from conscious status by fiat, as noted in Section 2. That axiom was introduced to eliminate the possibility of multiple consciousnesses cohabiting a single system in order to bring the results of the formalism into accord with phenomenology (and, presumably, parsimony); and now, upon trivializing reference to a point of view as a distinctive unifying factor, the unity of consciousness is formulated and axiomatized in a way that may well beg the question in favor of IIT and its specific formalism.

We conclude that short of a phenomenological account of this purported irreducibility (this unity or unities of consciousness, as the case may be) untainted by borrowings from the irreducibility requirements of the formalism, IIT fails to convincingly match phenomenology with formalism on both its “intrinsic perspective” and “irreducibility” construals of the phenomenological fact of the unity (or unities) of consciousness. In the case of the former construal, this is because the “intrinsic perspective” (and predecessors, viz., the “single point of view”) is not found in the IIT formalism per se and is trivialized when added to it ad hoc. In the later case, this is because it is simply unclear that the varieties of “unity” describable from the phenomenological point of view are all captured by the kind of irreducibility native to the IIT formalism, as the “Integration Axiom” implicitly assumes. Moreover, among these varieties, a strong argument can be made that figure-ground or Gestalt-type unities actually require a non-
trivial viewpoint structure for their adequate accounting. We turn then to the IIT treatment of
that other fundamental and undoubted attribute of consciousness, its “differentiation”.

One would think that the richly differentiated, composite nature of any given conscious
experience, most conspicuously in perceptual consciousness, would be a prime exhibit for a
robust notion of the differentiation inherent in consciousness. The closest we get to this is the
Composition Axiom (“Consciousness is compositional (structured): each experience consists of
multiple aspects in various combinations”, Oizumi et al. 2014, p. 2).\(^8\) This highly abstract
characterization, which is also convenient vis-à-vis the IIT formalism, is then tied to the unity of
consciousness via the “irreducibility” claim considered above, rather than to its concept of
differentiation. This is because IIT reserves the term ‘differentiation’ for something quite
different, namely the hypothetical “informativeness” of a given experience relative to all
possible conscious experiences. Thus: “the occurrence of a given conscious state implies an
extremely rapid selection among a repertoire of possible conscious states”; and “the occurrence
of one particular conscious state over billions of others therefore constitutes a correspondingly
large amount of information” (Tononi and Edelman 1998, p. 1846, cf. Edelman and Tononi 2000,
pp. 29-34). Or: “By definition, a highly conscious experience is a discrimination among trillions of
alternatives – it specifies that what is the case is this particular state of affairs, which differs
from a trillion other states of affairs in its own peculiar way” (Tononi, 2008 p. 240). By 2014, this
construal of differentiation was formulated as follows in the “Axiom of Information”:

\(^8\) Balduzzi and Tononi 2009 is an effort to develop a richer theory of qualitative and conceptual content in IIT
terms than what is offered in Oizumi et al. 2014, which refers to the former. We do not have space to examine this
paper in detail here. Suffice it to say that even if efficient network organization of the type IIT formalizes were
necessary, constitutively or causally, for content differentiation (in this richer sense), this does not mean that it would
be sufficient for consciousness (see Sections 4 and 6).
“Consciousness is informative: each experience differs in its particular way from other possible experiences. Thus, an experience of pure darkness is what it is by differing, in its particular way, from an immense number of other possible experiences. A small subset of these possible experiences includes, for example, all the frames of all possible movies” (Oizumi et al. 2014, pp. 2-3).

We would be lucky indeed were the informativeness of our momentary experiences concretely defined with respect to the universe of all possible experiences, and based on “discrimination among trillions of alternatives”! But such a claim conflates informativeness with information capacity, which are by no means equivalent concepts. The set of all distinct system states – in this case “all possible experiences” – is a measure of a system’s information capacity, in principle calculable from a knowledge of the number of elements that compose it and their causal connectivity. Informativeness, on the other hand, pertains to the significance of a given experience to the “haver” of that experience, and, loosely speaking, to the system itself in that given state or state transition. The extent to which that significance is a function of other experiences of which that system is capable is a matter of the extent to which the given state has direct or indirect access to or information about other states (experiences) of the system, an issue subject to innumerable factors regarding the organization of the system and its properties.

Fortunately, we need not enter into these complexities because the systems so far defined by the IIT formalism are memoryless (Balduzzi & Tononi 2008), precluding them from such access. The IIT systems so far studied dwell in a perpetual pristine present for which no content outside
the state it is in exists, precluding them from (meaningfully) registering either surprise, boredom, or interest.\(^9\) Informativeness, as characterized in the previous paragraph, is therefore not defined for such systems, though this does not mean that such a system is incapable of performing useful functions. So far bereft of a principled account of the informativeness of experience, IIT proceeds as if half a century of progress on the topic can safely be bracketed (Bindra 1959; Sokolov 1963; Sachs 1967; Hinton and Zemel 1994; Knill and Richards, eds. 1996; Rao and Ballard 1999; Merker 2004; Friston 2005; Rudrauf et al. 2017). ‘Differentiation’, as defined in IIT, thus joins ‘integration’ as another point where the phenomenology of consciousness and IIT formalisms fail to substantively and naturally match. But these two – integration and differentiation – are the principal points on which IIT rests its thesis of an identity between consciousness and integrated information.

One possible reason for the survival across successive versions of IIT of these questionable or question begging analogies and these mismatches between phenomenology and formalisms is the confinement of the mathematical apparatus of IIT to the information side of its attempted identification of integrated information with consciousness, with the arguable exception, just noted, that formulations from the information side make their way into the characterization of the phenomenology itself. While the information side is formulated in rigorous algorithmic terms, its phenomenological side is couched in heuristic terms, featuring intuitive illustrations.

---

\(^9\) Balduzzi and Tononi (2008, p. 18) promised to consider IIT systems that incorporate memory in a forthcoming paper but, to our knowledge, that paper was not published. But see Balduzzi 2011. Needless to say, equipping the system with memory capacity is but an elementary step towards defining informativeness, as the literature citations that follow in the text intimate.
and thought experiments, rather than primary phenomenology. This disparity in language and rigor vis-à-vis the two sides of the IIT identity thesis enables the kind of imprecise and outright mistaken bids to identify the two domains with one another reviewed in the foregoing.

This unsatisfactory state of the phenomenological side of IIT is not merely a consequence of the limitations under which the nascent science of consciousness is laboring at this early phase of its development (see Introduction). There is no dearth of structural, quantitative, and more rigorously framed qualitative phenomenological data by which to constrain and fortify the “consciousness side” of the *prima facie* distinct domains IIT seeks to equate. Perhaps most notably in the present context, more than a century and a half of psychophysics, much of whose material consists of quantified data about conscious phenomenology, lies untapped by IIT.

From the geometries of visual space (Wagner 2006) to the many dimensions along which sensory and cognitive contents of consciousness are articulated, psychophysics provides a wealth of empirical data based on quantitative methods and mathematical interpretation (Gescheider 1997; Price & Barrell 2012; Read 2015; Steingrimsson 2016; Rudrauf et al. 2017, Rudrauf et al. 2020a). To give but one example: since its very inception, psychophysics has been concerned with “the difference that makes a difference”, a concept prominently featured throughout IIT (e.g. Tononi & Edelman 1998, p. 1846; Oizumi et al. 2014, p. 9); yet on the phenomenology side of IIT, this construct remains completely abstract. Recourse to relevant

---

10 In Edelman and Tononi (2000, p. 217), the duo exhibited a common conflation of phenomenology with introspectionism; this has been subsequently, if implicitly, rectified in Oizumi et al. 2014, though this bit of meta-methodological improvement has not significantly improved the theory’s prospects on the phenomenological score, as noted.
psychophysics (for a mathematical treatment, see Iverson, 2006) might, in principle, allow quantification of the differences that actually make a difference phenomenologically to be weighed and tested against the formalisms on the algorithmic side of IIT. In practice, serious barriers might, of course, remain, given the general difficulties with the application of IIT to real systems, even on its own criteria, discussed in Section 2. But at the least a more thoroughgoing attempt to come to terms with the relevant psychophysics and phenomenology, especially as they bear on the “difference that makes a difference”, “unity” and “composition” issues would seem to be de rigueur. Even so, as we have argued for the latter two issues, serious attention to the phenomenology and psychophysics might well undermine the convenient formulations of their codifying axioms. Selecting the appropriate level of abstraction as well as the central phenomenological features around which to center one’s mathematical modeling are crucial. This is a non-trivial matter that requires close attention to the relevant psychophysics and phenomenology.

As things now stand, the under-constrained – even misconstrued – nature of the phenomenological side of the IIT identity thesis allows tropes and metaphors borrowed from its mathematical side to intrude into or even substitute for phenomenological data, as noted. However, if phenomenology is to offer support for the IIT thesis, it must be genuinely empirical phenomenology rather than heuristic arguments and thought experiments alone; and it must, above all, remain completely untainted by intrusions and borrowings from the side of the algorithmic formalism, at least when it is being appealed to as a primary and independent source of evidence for the identity claim. It is to little avail in this regard to fortify the
phenomenological side of the thesis by secondarily casting its claims and conjectures in axiomatic form, as has been attempted in the presentation of IIT since 2012. Its axioms and postulates provide, at worst, no more than a codification of the mismatches covered in the foregoing. As shown in a comprehensive review of the axiomatic format of IIT by Tim Bayne, its axioms either fail to qualify as “axiomatic” (in Tononi’s intended sense) or fail to provide substantive theoretical constraints (Bayne 2018). We refer the reader to Bayne’s treatment for details.

Here we merely supplement Bayne’s critique by noting that the difficulties encumbering the IIT “Axiom of Exclusion” are a consequence of not keeping the informational and the phenomenological sides of IIT methodically distinct. As we noted in Section 2, the ad hoc “Axiom of Exclusion” serves to eliminate the possibility of congeries of consciousnesses (in the plural) otherwise entailed by the IIT formalism, in violation of the phenomenological unity of consciousness under a point of view (as well as some general parsimony considerations). The problem solved in this arbitrary fashion would not exist if a unifying point of view could be derived from the IIT formalism as such, because there would then be only one such viewpoint, underwriting the unity of the whole with no room for subsets with potential viewpoints of their own. The violation of the phenomenological unity of consciousness thus removed, the need for an “Axiom of Exclusion” would disappear.

We end this Section by noting that even if we assume, for the sake of argument, that these mismatches and anomalies can somehow be removed from the theory, and it could be
established that consciousness *necessarily* requires a combination of integration with
differentiation of the very kind specified formally by IIT, this, absent further premises, would still
not *ipso facto* amount to a sufficient condition for consciousness and therefore still could not
serve to define it. All that can be concluded from the conjunction of integration with
differentiation as a *necessary* condition for consciousness is that in the absence of that
combination of attributes, there is no consciousness. At a minimum, the claim of sufficiency
would require independent evidence that all members of the great diversity of natural and
artificial systems that do in fact combine integration with differentiation (in the formal IIT sense)
are in fact conscious. No such evidence has ever been presented by Tononi and colleagues, a
matter to be further considered in Sections 5 and 6.

If, then, the IIT bid to identify integrated information with consciousness by the coincidence of
their defining principles fails, does not the co-variation of its proposed measure of
consciousness with a number of relevant empirical circumstances nevertheless support that
identity? We turn next to an examination of IIT's claims in that regard.

4. IIT in context, III: neuroscience

The intractability of computing Φ for a physical system such as a mammalian brain or its parts
prevents IIT from being subjected to direct empirical testing. To support its claims, it has relied
instead on proxy measures and substitutes for Φ, including extrapolation from toy systems. In
view of the problems encumbering such expedients already noted in the fourth and fifth paragraphs of Section 2, this is an approach beset with perils. A case in point is IIT’s frequently invoked claim about the contrasting contributions of the cerebral cortex and cerebellum to consciousness. Compared to the cerebral cortex, the cerebellar contribution in this regard is generally taken to be negligible (Romaniello and Borgatti 2013), a contrast which according to IIT is captured by contrasting levels of Φ generated by the two neural systems (Tononi and Edelman 1998; Tononi 2004, 2008, Oizumi et al. 2014).

That claim is not based on actual measurements of Φ for either cerebral cortex or cerebellum, however, but on assessing Φ on the causal connectivity of toy networks. Thus, a network of 12 elements grouped into 3 “modules” (Tononi 2008, Fig. 4B) or one of 6 elements, again in three modules (Oizumi et al. 2014, Fig. 17A), serve as stand-ins for the intricate connectivity of the approximately 100 billion neural elements of diverse types that make up the cerebellum (Andersen et al. 2003). Such a way of proceeding might make sense if the toy systems schematically approximated the well-known connectivity of the cerebellum, but this is not the case. The sole link between the IIT stand-in networks and the actual cerebellum is the assumption of a loosely defined “modularity”, implemented differently in the two publications just cited. But even that minimal commitment may be hazardous because the cerebellum is more than its cerebellar cortex, on whose organization the assumption of patch-like local modularity is based.
Embedded in the white-matter core of the cerebellum lie the cerebellar nuclei. These not only receive, via collaterals, extra-cerebellar information destined for the cerebellar cortex, but massive recurrent projections from the Purkinje cells of that cortex. Moreover, they issue sometimes widely branching axonal projections up to the cerebellar cortex itself. These branching axons span cerebellar zones as well as lobules, thus breaking the modularity of cerebellar organization (for a concise summary, see Haines and Dietrichs 2012; see also D’Mello et al. 2020). In some respects, this reciprocal (two-way) connectivity between cerebellar cortex and cerebellar nuclei is reminiscent of the organization of the thalamocortical complex. In this setting, the modularity-abrogating consequences of cerebellar nucleo-cortical connectivity make it impossible to say anything about cerebellar $\Phi$ levels without actually assessing them on credible models of its functional architecture, something never attempted in IIT. Short of that, claims relating $\Phi$ to the relative contributions of the cerebellum and the cerebral cortex to consciousness are moot, and cannot provide independent support for IIT’s identification of integrated information with consciousness.

As regards the cerebral cortex, given the origin of the measure $\Phi$ in the attempt to capture the way the cortex combines integration with differentiation (Tononi et al. 1994), there is no doubt that the functional architecture of the cerebral cortex (henceforth ‘cortex’) would yield high levels of $\Phi$, if that quantity could be measured empirically on cortical operations. That prospect has been brought nearer to realization by the Toker-Sommer advance reviewed in Section 2. As we saw there, $\Phi$ tracks the efficiency of global information transfer in differentiated causal networks (Toker and Sommer 2019). There is overwhelming support for this property of cortical

In Section 2, we noted the wide range of natural and artificial networks that employ efficient network design to carry out their functions. The common denominator of all these networks, from metabolic and gene regulation networks to electrical power grids and the cerebral cortex, is the need to optimize complex, multicomponent network interactions under temporal, spatial, and energetic efficiency constraints. The cortex maintains its wide-ranging inter-areal interactions of millions of components under all three of these constraints. It is the largest part of the mammalian brain, draped across its surface like a bark, giving it its name. Interactions across this extensive surface impose a distance cost on its operations (Cherniak 2012). Cortical activity bridges those distances at speeds some 30 million to 180 million times slower than electronic circuitry,\(^\text{11}\) adding a severe temporal constraint to its operations. These operations are also metabolically expensive (Laughlin 2001; Lennie 2003; Hyder et al. 2013; Niven 2016) and deliver their verdicts in real behavioral time some three to four times per second (the frequency of gaze movements, the leading edge of most behavior, Merker 2013b). The inevitable upshot of working under these severe spatial, temporal, and energetic constraints is

\(^{11}\) Cortical signal propagation (across all membrane components, axonal, synaptic, and dendritic) range from around 1 to 6 meters per second (data in Schmolesky et al. 1998; Pasqual-Leone et al. 2000). The speed of electrical transmission in copper or photons in fiber is about .6 of the speed of light, i.e., 180 million m/s, which at 2.4 m/s neural speed would give us a rate 75 million times slower than electronic transmission. Transistor state change delays are measured in negligible billionths or trillionths of a second.
that for any processes other than its strictly local ones, the cerebral cortex, whatever its functional roles, must be organized in efficient network fashion (cf. Modha and Singh 2010; Bullmore and Sporns 2012; Cherniak 2012; Markov et al. 2013; see also van Rossum et al. 2002), irrespective of how those roles may be related to consciousness.

The cortical network efficiency in question is one of global information transfer in flexible balance with local specializations (see Fig. 2A of van den Heuvel and Sporns, 2019). In such a network any node can be reached from any other node via but a few intervening nodes, though each of them is a member of specialized local or “regional” clusters. In the case of electrical power grids this enables swift redistribution of loads in the face of sudden fluctuations of diverse local or regional demands. In the cortical case, its efficient network organization extends up to its most global level, i.e., the cortex as a whole is an efficient network (Markov et al. 2013; van den Heuvel and Sporns, 2013). Might there be some generic cortical function that requires such a high Φ mode of organization at the global level, and therefore might account for this functional architecture without in any way warranting the consciousness conjecture advanced by IIT?

No cortical function lies nearer at hand in this regard than the process by which the cortex accomplishes what is arguably its cardinal function: efficient long-term storage of information (Braitenberg 1974; McClelland et al. 1995; Squire and Alvarez 1995; Merker 2004; Brunel 2016; Mansvelder et al. 2019). That process, which involves what is known as “systems consolidation” (reviewed by Miyashita 2004; Frankland and Bontempi 2005; Vinocur and Moscovitch 2011;
Dudai et al. 2015), is utterly unlike information storage in computer memory systems. For one thing, the storage process is not a one-off event, but extends over days, weeks, months and even years. For another, over its time course it transforms the stored information itself through changes in the nature and location of the synaptic modifications through which cortical storage takes place. And finally, in the course of these transformations it interacts with and affects earlier stored information through the distributed synaptic or connectivity changes that implement the consolidation process.

The open-ended nature of systems consolidation (and “reconsolidation”) is such that, in a sense, it never ends (Dudai 2012). This means that different kinds of information, acquired at different times and under different circumstances, are undergoing perpetual change through their interactions. For such a mode of information storage to be functional, rather than to degenerate into arbitrary distortions of its contents, something must ensure the coherence of the process as a whole. That is what the efficient network organization of the cortical system as a whole accomplishes. The network alterations that implement storage, whether they involve changed synaptic weights (Rogerson et al. 2014) or structural remodeling (Porazi and Mel, 2001; Chkloskii et al. 2004; Knoblauch et al. 2014), are driven by neural activity, and neural activity follows the connective structure of the network. When the latter is organized for efficient global information transfer (high Φ) any point in the network can reach any other with few intermediate steps. Activity anywhere is thus potentially capable of affecting activity anywhere else by local propagation.
The joint effect of all such propagating influences taken together constrains activity anywhere by the state of the system elsewhere and ultimately – because the network efficiency of the cortex is global – *by the state of the system as a whole*. The exponential distance rule of cortico-cortical connectivity (Ercsey-Ravasz et al. 2013) ensures that such influences are graded by synaptic distance. Synaptic distance in the cortex translates to functional dissimilarity (e.g. Passingham 2002). As a consequence, the cortical system, and the pattern of long-term change wrought by activity within it, is organized along contextual lines of relevance (see Merker 2004, for details). Such a contextually organized information storage capacity amounts to a coherent epistemic space. Colloquially we call it our “knowledge”, most of which lies outside the bounds of our consciousness at any given point in time.

Thus, the global efficiency of information transfer that provides the operative setting for “systems consolidation” rescues cortical long-term memory from its potentially self-defeating interactive malleability by imposing contextual coherence on its organization. When disposed in countercurrent fashion, as the cerebral cortex manifestly is (Markov et al. 2013),\(^1\) such a network becomes a powerful engine for the inductive accumulation and storage of what, in Bayesian terms, amounts to the priors needed for the feats of inductive inference that have preoccupied cortical theorists over the past three decades (Mumford 1992; Ullman 1995; Dayan...)

\(^1\) Globally this network implements a cortex-wide double and staggered countercurrent mechanism in the form of a “core-periphery” network with a dense core “bow tie” topology (Markov et al. 2013; see Doyle and Csete 2011, for other examples of “bow tie” network structure in biology and technology, and Vitali et al. 2011 for an example from economics). In the cortical case its path-length minimization is near optimal (Markov et al. 2013), and its countercurrent organization (which “turns over” in the hippocampus, Merker 2004) is crucial for the generative and predictive operations central to its probabilistic mode of operation. There is not a hint of the cortical countercurrent in IIT diagrams, nor do they feature even a basic implementation of thalamocortical architecture. Their resemblance to anything neural is limited to the combination of differentiation with integration central to IIT, whose toy network diagrams are didactic devices useful for teaching the concepts of the IIT formalism in simplified form. They are not derived from the way actual brains are organized.
and Hinton 1996; Knill and Richards 1996; Rao and Ballard 1999; Friston 2002, 2005; Merker, 2004; Hinton, 2007). Add to this the evolutionary expandability and functional resilience to path loss of the cortical network (Markov et al. 2013; see also Doyle and Csete 2011; Rudrauf 2014), and it becomes clear that there are multiple compelling and convergent reasons, irrespective of any hypothetical relation to consciousness, for the functional architecture of the cortex to exhibit efficient network organization and correspondingly high $\Phi$. Short of evidence that it is consciousness specifically that mandates a cortical organization optimized for global information transfer, measures of cortical $\Phi$ provide no support (neither *prima facie* nor *secunda facie*) for IIT’s identification of $\Phi$ with consciousness.

But has not IIT provided exactly such evidence, that is, evidence that it is consciousness specifically that is served by the efficient network organization of the cortex, by showing that $\Phi$ varies in appropriately graded fashion with departures from wakefulness such as dreaming and dreamless sleep, anesthesia, seizures, and coma (e.g. Massimini et al. 2005; Ferrarelli et al. 2010; Casali et al. 2013)? Beside the fact that it is stand-ins for $\Phi$ and not $\Phi$ that have been measured in these circumstances, the problem here is that these are the very conditions under which the performance requirements for cortical operations dependent on efficient network interactions are relaxed or absent. When there is no need for targeting gaze movements, say, there is no need for efficient fronto-parietal network interactions for gaze control, and similarly for all other extra-local cortical traffic, up to the global level of systems consolidation reviewed in the foregoing. The brain has no reason to sustain these interactions when they are not needed. In fact, going into the various states of less than full wakefulness is how the cortex
behaves when the need for its prime service of efficient network organization is lessened or absent! The demand for these wide-ranging interactions is lessened across the stages of sleep, and altogether absent in anesthesia or coma (see, e.g. Mashour 2013). So it is hardly surprising to see Φ reflect the extent to which they are in fact engaged.

Stand-in Φ measures, in reflecting the extent to which cortical operations served by network efficiency are in fact operationally patent and engaged, accordingly provide potentially useful indices of the extent to which the cortex is working normally and making its particular contribution to the brain's over-all functional division of labor. To go from such empirical reflections of cortical engagement generally to identifying Φ with consciousness specifically would require experimental dissociation of the consciousness variable itself from the operations that are dependent on efficient cortical information transfer and crowned by systems consolidation, a tall order unlikely to be achieved in the foreseeable future.

Finally, we note that contrary to IIT’s predictions, surgical severance of the principal connection between the two cerebral hemispheres (the corpus callosum) appears to leave the unity of consciousness intact (Pinto et al. 2017). Roger Sperry (1968, 1984) and Michael Gazzaniga (1967) based their conclusion that split brain patients have two separate consciousnesses, one for each hemisphere, on partly inconsistent, non-quantitative assessment of the laterality of patient response mode (left versus right hand, or verbal response, the latter left-lateralized) relative to which hemisphere “saw” the laterally presented visual stimuli. As far as corticocortical anatomy after callosotomy is concerned, a visual stimulus which because of its
presentation to the right of fixation is represented in the left hemisphere should be reportable both through the right hand (controlled by the left hemisphere) and verbally (left lateralized language), while one presented to the left of fixation should be reportable only through the left hand and not verbally. According to Sperry and Gazzaniga, patient performance conformed to this pattern, yielding their conclusion that two separate conscious agents were present in these patients. Careful quantitative assessment of response mode in split brain patients by Pinto and colleagues confirms that their visual fields are indeed split between the hemispheres as reported by Sperry, but their response mode is not. Contrary to the original claims and lingering textbook accounts, they act as single conscious agents: they are capable of responding with either hand as well as verbally to stimuli whether they are presented to the right or to the left of fixation (Pinto et al. 2017; de Haan et al. 2020).

Knowing exactly what these new findings imply regarding neural mechanisms of consciousness more generally still depends on answers to some interpretive issues raised by de Haan and colleagues; but one distinct possibility is that no theory of consciousness couched in terms of cortical circumstances alone can be complete. Here we only note that in terms of IIT, the severance of the principal conduit of communication between the two cerebral hemispheres is a “cruel” cut indeed, with a correspondingly drastic impact on information integration. Thus, according to IIT, the split-brain syndrome yields a measure of integration for each hemisphere that is larger than for the two hemispheres taken together. Under its principle of exclusion, and assuming rather boldly no other substantive (e.g., subcortical) paths of integration, this should result in two independent conscious systems rather than one conscious agent in a split brain
patient (Tononi 2005), as originally claimed by Sperry and Gazzaniga but contradicted by the quantitative assessment performed by Pinto and colleagues.

In sum, we submit that IIT is a theory of integrated information of the kind defined by what we have called efficient networks rather than of consciousness. It is the cortical need to perform complex multicomponent operations under temporal, spatial, and metabolic constraints at a global level that accounts for its formal kinship with systems as far afield as gene expression networks, large-scale electrical power grids, social networks, and arrays of exclusive-or gates arrayed in expander graph fashion. To identify integrated information with consciousness over and above the interpretation offered by network efficiency amounts to an easily avoided overreach. Because Φ in fact reflects the efficiency of global information transfer, it is the efficient network organization of the working cortex that gives it its high Φ score (and similarly for the non-neural systems), irrespective of whether consciousness is one of the functions served by that organization. From the point of view of neuroscience, there is thus no reason to identify Φ with consciousness specifically, but there is every reason to expect Φ to reflect the state of cortical operations. IIT insists, however, on identifying Φ with one of the functions served by cortical network efficiency (the latter presumably being a requirement for rich and sophisticated contents of consciousness as we know it), rather than with the network efficiency Φ actually measures. This amounts to a functional misassignment – the mistaken identity of our title – whose consequences have led IIT from ordinary science towards the orbit of panpsychist ideation, a matter to be considered next.

13 Needless to say, the upshot of cortical operations does end up shaping the rich contents of our consciousness, though how and where is still an open question (Merker 2012, 2013b).
5. From functional misassignment to pan(proto)psychism

In Section 2, we saw how, at the beginning of the development of their formal measure of integrated information, Tononi and Edelman expected it to apply to a wide range of complex systems, from networks of gene regulation to communication networks (Tononi and Edelman 1994; see also Marshall et al. 2017). At the time, this expectation was motivated by considerations of causal network complexity alone, rather than by any inherent connection between causal complexity (integrated information) and consciousness. Section 3 reviewed the failure of IIT to make good its direct identification of integrated information with consciousness in both its “integration” and its “differentiation” aspects. And in Section 4 we saw that the cerebral cortex, given the constraints that encumber its operations, must exhibit globally efficient network organization to accomplish its demanding functions, and its crucial information storage function in particular, however those functions might relate or contribute to consciousness.

Recall the fact with which we concluded Section 3, namely, that even if consciousness as we know it necessarily features a combination of integration with differentiation, it does not follow that wherever that combination is present, there is consciousness. If it is only a necessary condition for consciousness, logic dictates simply that wherever that combination is missing there is no consciousness, not that wherever it is present, there also is consciousness. It is
fortunate, therefore, that the labors of network science have yielded a truly diverse set of
natural and artificial systems that are well characterized in terms of their high capacity for
efficient global information transfer in differentiated networks, the very network property
measured by Φ. As we have noted, that set includes metabolic and gene regulation networks,
social networks, communication, transportation, and power distribution networks (including the
internet), a variety of electronic circuit board designs (e.g., “expander graph”), and neural and
semantic networks (Boccaletti et al. 2006; Aaronson 2014). In each of these cases we know that
their capacity for efficient global information transfer yields substantial Φ scores (Toker and
Sommer, 2019).

The two claims bearing on the empirical distribution of Φ across such systems are these:

1. The efficiency of global information transfer in differentiated networks is the
functional factor that co-varies with Φ, with no need to invoke consciousness merely on
the basis of its presence.

2. Consciousness is the factor that co-varies with Φ, because consciousness is integrated
information, which is what Φ measures.

In the case of Claim 1, cast in terms of network efficiency as defined here, we know for a fact
that every one of the diverse systems just enumerated is an efficient network, yielding high Φ.
For Claim 2, the claim central to IIT, we have direct evidence of consciousness for only a single
one of these system kinds, namely the human brain and, based on multiple analogies (e.g.
Butler 2008), circumstantial evidence for some other brains, but not for any other natural or

artificial systems such as power grids and gene expression networks. As we know, IIT
nevertheless insists on identifying integrated information with consciousness (Tononi 2004,

Pending independent evidence of consciousness across the diverse set of natural and artificial
objects that score high on Φ, Claim 2 seems gratuitous, quite apart from the many grounds for
doubting the central claim of IIT we have reviewed in Sections 2 through 4. Yet IIT, in the
absence of such evidence, is committed to Claim 2, and accordingly attributes consciousness to
gene regulation networks, large scale electrical power grids, etc. solely on the basis of their
substantial Φ scores.

With tentative beginnings in 2008 (Tononi, 2008, p. 236), Tononi and colleagues have found a
potential rationale for that commitment in the realm of panpsychist ideation: “...in line with the
central intuitions of panpsychism, the integrated information theory treats consciousness as an
intrinsic, fundamental property of reality” (Tononi and Koch 2015, p. 11). According to
panpsychism, “psyche”, mind, or consciousness is not uniquely associated with brains or brain-
like objects, but enjoys a much wider distribution, as required by IIT.

IIT cannot, however, fully embrace panpsychism in its classical form of postulating mind
“everywhere and all the way down”, i.e., the stance that psyche, mind, or consciousness is a
fundamental and ubiquitous feature of reality and that everything therefore is endowed with
consciousness or consciousness-like properties. This will not do for IIT because if everything is
endowed with consciousness, IIT is trivialized as a constitutive theory of consciousness. The entire computational apparatus supporting the $\Phi$ measure would be gratuitous with respect to the constitutive issue of consciousness per se, since many systems scoring zero on that measure would nevertheless possess consciousness, according to panpsychism as classically conceived. IIT accordingly eschews panpsychism in this sense: “Unlike panpsychism, however, IIT clearly implies that not everything is conscious” (ibid.). The key to this “partial panpsychism”, to put it oxymoronically, is supplied by the cryptic statement that follows to the effect that “IIT offers a solution to several of the conceptual obstacles that panpsychists never properly resolved, like the problem of aggregates (or combination problem)” (ibid., citing Chalmers 2016b, and James 1890 for the “combination problem”).

Tononi and Koch do not spell out this promised solution, yet their reference to Chalmers and the “combination problem” hints at what they have in mind. The metaphysics of consciousness Chalmers (not completely unfavorably) considers in this connection does not postulate mind, consciousness, or phenomenality “all the way down”, but only “protophenomenality” (Chalmers 2016a). Somewhere between ubiquitous protophenomenality and consciousness proper there would accordingly have to be a transition from one to the other. Protominds are simple and numerous, while conscious minds like ours are complex and less numerous. Could it be that IIT, having started as an attempt to understand the functional organization of the cortex, has
uncovered the conditions under which protophenomenality combines or aggregates in the transition to consciousness proper?\textsuperscript{14}

If that were true, conscious power grids and gene regulation networks would cease to be compelling counterexamples to IIT. Moreover, the prospect of a complete integration of panprotopsychism with IIT holds out the tantalizing possibility of fomenting a scientific revolution by showing that science as we know it has misconstrued the boundary conditions for the genesis of consciousness by missing a critical factor at the very foundations of our universe, as Chalmers has maintained elsewhere on speculative grounds (Chalmers 1995, 1996). Chalmers refers to Tononi and IIT, albeit non-committally, in his paper on the combination problem (Chalmers 2016b), while Tononi and Koch, as we saw, cite that paper of Chalmers on the same topic. Nor are they alone in pursuing the prospect of non-trivially relating IIT to some version of panpsychism (see, e.g., Mørch, 2018). Note, however, that the special and irreducible “mindish” ingredient that panprotopsychism would postulate, motivated ultimately by an entirely different set of (dubious) arguments (see, e.g., Williford 2020), is what would, on such a theory, be providing the rock-bottom explanation of phenomenal consciousness as such, and not IIT, which, in that case, would at best provide a story about how conscious minds like ours (and like those of electricity grids, etc.!) come about via the sort of “combination” it describes.

\textsuperscript{14} Note that, depending on how one conceives of protominds or protophenomenality, there may be absolutely no conceptual advance here over the analogous problem of how physical components combine to realize conscious minds. In that sense, any “alliance” between even panprotopsychism and IIT may be devoid of any substantive empirical or even theoretical content. A panprotopsychist IIT would then, after much ado, come down to a mere terminological variant of a physicalist version of IIT. Thanks to Miguel Ángel Sebastián for discussions on this point.
Given the mismatches detailed in Section 3 and the ready availability of an evident account of high Φ scores (in the brain and elsewhere) provided by network science—and in fact by IIT itself, considering what its formalism actually measures—no reason to abandon science as we know it for panpsychist or panprotopsychist metaphysics can plausibly be drawn from IIT. If, therefore, as the sum total of our analysis suggests, IIT can be safely set aside as a failed bid for a scientific account of consciousness, what are the prospects, in the third decade of the twenty-first century, for arriving at a better account? And what principles are available to guide and constrain the search for it?

6. What then is consciousness, if not integrated information?

In the foregoing we have examined a rather comprehensive set of aspects of IIT vis-à-vis its claim to offer a theory of consciousness. For each aspect examined, we have found IIT to be encumbered by serious shortcomings. In the aggregate, these shortcomings converge on the conclusion that IIT is not only under-constrained as a theory of consciousness, but that it is a theory of something other than consciousness, namely of that which, for short, we have called network efficiency. As we have been at pains to point out, the powers of efficient networks are such that numerous natural and artificial systems have availed themselves of their benefits for reasons that can be specified without warranting any reference to consciousness whatsoever. Nevertheless, IIT, as a theory of consciousness, insists on identifying the integrated information defined by global information transfer in differentiated networks with consciousness itself.
(Tononi 2004; Oizumi et al. 2014). Throughout the development of IIT this identity claim has been supported by a small set of intuitive illustrations and thought experiments. These intuitive commitments were eventually clothed in spuriously axiomatic form (Tononi 2012; Bayne 2018), though their content predates IIT itself (Tononi & Edelman 1998, 2000).

Such a less than rigorous way of proceeding is not uncommon in the early stages of a nascent discipline. As noted in our Introduction, by way of quoting Thomas Kuhn, the safeguards of generally shared and agreed-upon norms, definitions, and first principles are not available to the pioneers laboring in the pre-history of a science. However, when conjectural commitments made in such a situation occasion attribution of consciousness to power grids and gene expression networks and associated excursions into the ideational domain of panpsychism (Tononi and Koch 2015), they deserve scrutiny, as attempted in Sections 3 and 5. We found the multiple mismatches between the two sides of the IIT identity claim, the concentration of actual formalism to one side of it, and the questionable conceptual borrowing between its two sides to provide ample grounds for rejecting that claim and with it IIT’s claim to have constitutively identified what consciousness in fact is.

A lesser possibility nevertheless remains, namely that integrated information of the kind defined by efficient networks might be a necessary, though not sufficient, condition for a conscious mode of operation, i.e., that consciousness is present only in systems organized in efficient network fashion. That possibility is, of course, not eliminated by rejecting the identity claim, but it also evidently does not follow merely from the fact that creatures like us, in possession of
neural mechanisms capable of efficient global information transfer, also are conscious. The existence of natural and artificial systems whose efficient network organization can exhaustively be accounted for without warranting any reference to consciousness means that some evidence apart from a capacity for efficient global information transfer is needed to motivate a linkage between the two in any particular case. Once the identity claim has been rejected, showing that global information transfer in differentiated networks is a necessary condition for consciousness will ultimately hinge on some characteristic or attribute of consciousness itself, some defining property of consciousness, that can only be realized in networks of that type.

That human consciousness features information generated by neural operations dependent on efficient global information transfer – integrated information in IIT terms – is beyond question. We are most assuredly the beneficiaries of such information, as outlined in Section 4 in connection with the logistics of cortical information storage. That, however, does not tell us either why or how such information enters consciousness. For one thing, much of human information processing appears to take place without consciousness (for which see Velmans 1991; Kihlstrom 1996; Doyle and Csete 2011; Merker 2012; van Gaal and Lamme 2012). For another, there are no a priori grounds for ruling out the existence of species who lack differentiated networks with efficient global information transfer but are nevertheless possessed of consciousness. What is needed to settle such questions is a sufficiently precise specification of the nature and structure of consciousness to be able to determine under what conditions consciousness might exist.
But isn’t such a specification exactly what the quest for neural correlates of consciousness (NCC) is designed to supply, and – for good measure – in terms of actual neural mechanisms? Defined as “the minimal set of neuronal events jointly sufficient for a specific conscious percept” or “specific phenomenal conscious state” (Koch, 2004, pp. 104, 304), the NCC have been pursued by intensive experimentation ever since a paradigmatic example employing binocular rivalry was introduced by Logothetis and Schall in 1989 (Logothetis & Schall 1989). However, decades of clever experimentation in many laboratories notwithstanding, there is still not a single generally agreed upon neural correlate of consciousness on record. Instead, the quest has settled into two incompatible camps: the “report” camp on the one hand (for which see Dehaene and Changeux 2011; van Vugt et al. 2018) and the “no report” camp on the other (see Lamme 2006; Tsuchiya et al. 2015). They continue to produce convincing empirical evidence for the soundness of their respective positions only to be met before long by a counter-demonstration by the other camp.

The seeds of this impasse were sown by defining the NCC in terms of correlates on the one hand and neural conditions “sufficient for a specific conscious percept” on the other. Correlates can be quite remote, after all: the tangible correlate of internal combustion engines represented by Sri Lankan rubber plantations tells us little of interest regarding such engines. The apparent precision of a delimitation to a "specific conscious percept" is also problematic. By studying the conditions under which a specific percept enters consciousness one is actually studying the conditions under which that particular percept manages to intrude on an already perfectly conscious condition, albeit featuring some other percepts than the experimentally relevant one.
Depending on what one chooses to include in one's criteria for such intrusion or capture – for example, whether the ability to report its occurrence is included or not – one ends up with different minimal conditions and thus with different neural correlates.

Given this, it is hardly surprising that competent investigators have ended up drawing different conclusions, and have coalesced into opposing camps. As Ned Block has argued, the bone of contention is not so much first-order empirical evidence as a “constitutive” issue, namely whether the cognitive underpinnings of being able to report on (Block 2008), or even think about (Block 2019) experience form a proper part of what it is to be conscious in the first place. Determining whether those cognitive underpinnings do form such a proper part turns on determining what consciousness in fact is in its own right. To know that, one needs to know, not "the minimal set of neuronal events jointly sufficient for a specific conscious percept", but "the minimal set of neuronal events jointly sufficient for consciousness as such", a very different matter. Indeed, experiments designed in reliance on a definition of consciousness not accepted by the other camp cannot resolve the impasse between them. Resolving it requires an answer to the constitutive question that both can, at least provisionally, accept; and the question is where to find such an answer?

IIT cannot supply the criterion for consciousness needed to adjudicate between the two NCC camps, though one of them may consider the kinship between its commitment to a neural feedforward-feedback loop as key to conscious perception and the "recurrence" central to the IIT definition of a positive Φ to be grounds for such a hope (Lamme 2010; Tsuchiya et al. 2015).
As we have been at pains to argue, however, the IIT failure to support its identity claim means that it cannot tell us what consciousness as such is, and therefore cannot provide the constitutive support the no-report camp seeks. Till we have an answer to that question, we also cannot know whether efficient global information transfer in differentiated networks might be a necessary, though not sufficient, condition for consciousness, and thus whether conscious creatures without such neural circuitry might exist.

We appear to have arrived, in other words, at a point where it is no longer possible to do as Francis Crick and Christof Koch did in helping to launch the recent wave of concern with the topic of consciousness in neuroscience with their influential paper “Towards a neurobiological theory of consciousness” (Crick and Koch, 1990). Cognizant of the early stage of the project outlined in that paper they adopted an approach of circumspection and minimal initial commitments. Topics they deliberately and explicitly “left on one side” included “any attempt at a formal definition” of “what consciousness is”, as well as “arguments about what consciousness is for” (Crick and Koch, 1990, p. 264).

Such initial reticence when it comes to defining even the very topic of inquiry leaves investigators free to pursue intuitive hunches of the kind reflected in the diversity of proposals for neuronal accounts of consciousness listed in our Introduction. By the same token this way of proceeding is fraught with hazard, because, without being anchored to a conception of their subject matter robust enough to eventually lead to the target, hunches will be pursued in vain. Ultimately, of course, and supposing things go as they ought, only the approach that gets its
subject matter right will be crowned with success, and that requires coming to grips with the very questions Crick and Koch wished to keep in abeyance. The NCC impasse and theoretical commitments that, for example, invite excursions into the ideational domain of panpsychism seem to be telling us that those questions require better answers.

If so, constitutive and related theoretical questions can no longer be left off the programmatic agenda of the neuroscience of consciousness. We need to come to grips with the essential question of what defines a condition of consciousness in the first place, the formal requirements, ultimately in mathematical terms, for specifying such a condition, and, finally, how such a condition, once specified, might be implemented neurally. With questions such as these we find ourselves traveling down the David Marr programmatic hierarchy (Marr & Poggio 1977) – from formulating the problem, to solving it algorithmically, to implementing it in a neural or other medium. One does not have to treat that program as a prescription for sequential implementation in order to recognize its hint that addressing the top, constitutive, level cannot be postponed forever.

We are of course not about to present an answer to the constitutive challenge in this paper, whose main topic is IIT’s failure to meet that challenge in its account of what consciousness in fact is. However, some of the specifics of that failure, dealt with at length in the foregoing, do turn on issues that have to be resolved for continued progress towards a future conceptually and empirically robust theory of consciousness. As an example of one significant such issue, we return to the question of the unity and mode of integration of consciousness under a point of
view, discussed in Section 3. The problems encumbering the IIT account of the unity of consciousness discussed there turn on IIT having trivialized reference to unity under a point of view in accounting for this central feature of the phenomenology of consciousness (Oizumi et al. 2014). Its case for unity rests on intuitions and thought experiments subsumed under the Axiom of Exclusion, which was introduced, in a potentially circular way, to save the IIT formalism from entailing the existence of multiple consciousnesses in a single system. All this can be avoided by taking the unity of consciousness under a point of view, and irrespective of its contents (visual, auditory, bodily, and so on), as a primary and non-trivial phenomenological datum. In doing so, one can look to psychophysics rather than thought experiments to rigorously identify and characterize many of the phenomenological facts.

In Section 3 we insisted that psychophysics provides a rich source of empirical evidence bearing on the phenomenology of consciousness, and this is a case in point. Psychophysics has long since worked out for us important features of the viewpoint we occupy as first-person observers, at least as far as the modalities of vision and audition are concerned. Ewald Hering’s 1879 “law of visual direction” allows the visual egocenter (the technical term employed in psychophysics) or viewpoint to be experimentally pinpointed with millimeter precision on an individual basis (Hering 1879; see also Roelofs 1959; Howard & Templeton 1966; Barbeito & Ono 1979). The approach has since been extended, albeit with far greater experimental difficulty and lesser precision, to audition (Cox, 1999; Neelon et al. 2004; Sukemiya et al. 2008). Thus determined, the visual egocenter is found to be, first of all, single or cyclopean (not a foregone conclusion given that we have two eyes) and located in the midsagittal plane, just
behind the midpoint between the eyes, some 4–5 cm behind the bridge of the nose. Such a location, as in the left panel of Fig. 1, agrees well with our vaguer intuitive sense of “where we are located” when assessed empirically (see Limanowski & Hecht 2011; Starmans & Bloom 2012).

Figure 1. Point of view and egocenter. Left panel: The visual egocenter of the first author assessed by the method of Howard and Templeton plotted on a magnetic resonance image of the skull at eye level. Right panel: Ernst Mach's famous drawing of the phenomenology of a monocular view of his studio from his egocenter. The dark fringe of his eyebrow, the silhouette of his nose, and the edge of his moustache frame the view. These close-range details are typically not as crisply defined in visual experience as in the drawing, and in a full cyclopean view with both eyes open the nose typically disappears from view (see Harding 1961, for a detailed first-person account). Adaptations of Mach's drawing appear in Tononi & Koch 2015, an apparent allusion to the subjective perspective under a point of view that is no longer featured in the presentation of IIT. The original drawing appears as Figure 1 in Mach (1897, p. 16). It is in the public domain, and is reproduced here in a digitally retouched version, courtesy of Wikimedia (http://commons.wikimedia.org/wiki/File:Ernst_Mach_Innenperspektive.png).
The habitual position of the viewpoint/egocenter is acquired and calibrated on the basis of experience, as shown by the difference of its location in cases of congenital blindness, blindness acquired later in life, and the normally sighted (Sukemiya et al. 2008; see also Cattaneo & Vecchi 2011, pp. 21-23). In the congenitally blind, the auditory egocenter is located close to the center of rotation of the head. With visual experience it is “pulled” towards the visual egocenter, while presumably also exerting some “pull” on the latter, so that in the normally sighted the two essentially share a location a bit behind the line joining the center of rotation of the eyes, as in the left panel of Fig. 1. The egocenter can also, under certain conditions, such as in out-of-body experiences, be dislodged from its location behind the bridge of the nose to be suspended in the space of the experienced world apart from the experienced body (Blanke et al. 2004; Metzinger 2005; Ehrsson 2007; Rudrauf et al. 2017), implying that the egocenter is a product of a dynamic process, integrating multiple sensory modalities, whose parameters are not innately constrained, in keeping with what is said in footnote 13.

Besides the fact of its existence, the empirically determined location of the perceptual egocenter harbors additional implications for consciousness theory. It confronts us with the stark contrast between the first-person (phenomenological) deliverances of consciousness and the third-person description of the relevant facts. Confining ourselves to vision only for the moment: phenomenologically we face the world from our egocenter/viewpoint along straight

---

15 One of us has proposed that this first-person egocenter, or viewpoint, is only the innermost one of the three most global invariants that the brain’s probabilistic inferential labors extracts from the movement-contingent clusters of correlated (and uncorrelated) variances playing across its sensory systems in the course of our mobile engagement with the world (Merker 2013b, p. 13; see also Merker 2012, p. 54, and Philipona et al., 2003, 2004; Dean et al. 2004). Thus interpreted, it supplies the necessary pivot around which the compensatory coordinate transformations between the other two super-clusters, “body” and “world”, can be managed most economically during movement.
and uninterrupted lines of sight. They pass from this single viewpoint inside our skull to the world in front of us through an open (empty) cyclopean aperture occupying the upper face region of this experienced body of ours. Yet, as we know, this location – some 4 to 5 cm behind the bridge of the nose – is, objectively speaking, surrounded on all sides by opaque biological tissues (see Fig. 1, left panel). How, then, is it possible to have unobstructed lines of sight into the world from a location inside our heads that is surrounded on all sides by opaque tissues?

That is one of the more acute questions the psychophysics of spatial phenomenology poses to consciousness theory. In brief, our experienced head is the head of the neural virtual body included in the brain’s neural simulation of reality, for which arrangements are possible that are not realizable for the physical head itself (see Merker 2012, pp. 53, 55 and Merker 2013a, pp. 26–27 for details; see also Rudrauf, 2014, and Rudrauf et al. 2017, Williford et al., 2018 for the “virtual” nature of such models). We bring the matter up here only to highlight the way in which a viewpoint-based approach elicits the kinds of questions that need to be answered by any theory of consciousness that aspires to paradigmatic status.

A number of investigators have included reference to a point of view and the associated perspectival attributes of the phenomenology of consciousness in their models of and thinking about consciousness. We already saw Tononi and Edelman (1998) cite William James (1890) to this effect. Other more recent examples include, but are not limited to, (famously) Thomas Nagel (1974), Arnold Trehub (1977, 2007, 2013), Max Velmans (1990), Alex Green (2002), Steven Lehar (2003), Merker (2012), Rudrauf et al. (2017) and Peter Godfrey-Smith (2019).
However diverse the theories espoused by these authors might be in other respects, they accord a central place to a point of view and the perspectivalism that goes with it in characterizing the first-person phenomenology of consciousness. Though some of these treatments are formulated primarily in visual terms, and the term viewpoint itself is a visual metaphor, the significance of the construct is by no means limited to vision, as the already cited studies of the auditory egocenter, and its normal coincidence with the visual one, indicate.

Rather, phenomenologically, all the senses are integrated into a supramodal virtual space that encompasses both body and world, and is perspectivally organized (Rudrauf et al, 2017), even in the blind (Heller and Kennedy 1990; Tinti et al. 2018).

In fact, a "point of view" (“egocenter”) appears to be a general, formal and necessary constituent of the integration of multimodal information into a supramodal 3-dimensional space organized in a perspectival manner that frames human phenomenal consciousness, as recently formalized in the Projective Consciousness Model (Rudrauf et al., 2017, 2020a, 2020b, Williford et al. 2018). It casts this space as a 3-dimensional projective space subject to projective transformations in accordance with the formalisms of projective geometry. Such projective spaces in three dimensions are governed by the group of projective transformations in four dimensions, the so called Projective General Linear Group PGL(4). A simpler example, to get the intuitive idea of the role of the point of view in framing information in projective geometry, is the case of projective spaces in two dimensions P(2), with which we are acquainted through pictures of 3-dimensional objects (perspective paintings, photographs). The patterns that compose the picture are 2-dimensional objects containing a projection of a surrounding 3-
dimensional space, based on an origin (the origin of projection). This origin must reside outside of the picture plane itself, along a third dimension in a 3-dimensional vector space V(3). In the 2-dimensional case, that vector space can coincide with the real physical 3-dimensional space, and the origin can be located in that latter space. In a camera, for example, that origin would correspond to the optical center of the camera through which all rays of light pass after crossing the lens and before reaching the sensor or film plane. In the projective model the plane of projection is normally in front of the origin, however, as the canvas is in front of the eyes of a painter. The picture plane itself is, in any case, excluded as a possible location for the origin by geometric necessity.

In the 3-dimensional case, which is the one relevant for human consciousness, the origin of the projection must also reside outside of the 3-dimensional projective space along a fourth spatial dimension in a 4-dimensional vector space V(4), a dimension obviously not available in our experienced space of three dimensions. This mathematical constraint has two significant consequences for consciousness theory. For one thing, it accounts for the singularity and uniqueness of the subjective origin of consciousness: there is only one 4-dimensional origin \{0\}, from which all possible perspectives on the 3-dimensional ambient space are taken through PGL(4). For another, it explains why this origin or viewpoint cannot appear in perception or more generally be simply perceptually objectified as an object in the ambient 3-dimensional space, not even through the imagination, which is governed by the same constraints according to the model. Thus, the projective geometrical formalism automatically excludes the viewpoint from what is being “viewed”, in good agreement with the phenomenology.
This fourth spatial dimension, and more generally the 3-dimensional subjective space in perspective which depends on it, as entailed by the projective geometrical model, is a necessary, though not a sufficient, condition on any theory of consciousness that would do justice to the phenomenology of human consciousness as we know it. Completing the steps down the Marr program from this formal specification of the projective geometry (i.e., the algorithmic level) to its concrete implementation in neural circuitry is a further, as yet outstanding, matter. At first blush it might seem as if the physical deployment of neurons in three-dimensional space renders the neural implementation of a fourth spatial dimension problematic, but this is not so. Brain evolution has repeatedly solved problems of far higher dimensionality than the four dimensions required here, olfaction being a prime case in point. The similarity spaces for mapping olfactory stimuli are irreducibly high-dimensional (Cleland 2008), and though the exact number of dimensions involved is still being studied, they amount to many multiples of four (Mamlouk & Martinetz 2004; Magnasco, Keller & Vosshall 2015).

No issue of fundamental principle therefore stands in the way of searching for a neural topology that implements the viewpoint relation inherent in the perspectively organized supramodal space of human conscious phenomenology. Though still a matter for the future, we note that success in that endeavor would provide a neuronal as well as more conceptually satisfying theoretical point of attack for the elucidation of the constitutive aspects of consciousness, rather than mere correlates. That, in light of what has been discussed in this Section, would obviously be a high priority desideratum for progress towards a credible paradigm of
consciousness. It would also provide a fulcrum for dealing with the perspectival nature of consciousness and its realization of a “point of view”, an issue which Thomas Nagel in his famous paper suggested might pose an inherent obstacle to a scientific account of consciousness (Nagel 1974) but which, in light of the above, may instead provide, with some irony, the very key and entry point to such an account.

We end on this optimistic note regarding the potential for progress latent in re-orienting the neuroscience of consciousness around constitutive questions, fully cognizant of the thorny issues that must be resolved in order to make such an endeavor a viable one. This is not the place to expand on the many ramifications and challenges of moving constitutive questions to the top of the consciousness agenda in neuroscience. We have raised the question here because it was forced upon us by the case of mistaken identity that supplies the subject matter of a major programmatic bid for a scientific account of consciousness. That error turns on a constitutive question concerning what consciousness in fact is and how it is structured, attention to which may also resolve the stalemate afflicting the supposedly more neutral, empirical search for the neural correlates of consciousness. With such a focus, the neuroscience of consciousness, and consciousness theory more generally, might be better equipped to tackle its far from finished labors of characterizing the nature of consciousness and its actual implementation in terms of formal and neural mechanisms, a prospect which the issues raised in this article are meant to further.

It has, of course, long been at the top of the agenda in philosophy; but we have in mind a much more constrained approach to the constitutive issues (see, e.g., Williford et al. 2018, esp. pp. 2-10).
Conflict of interest declaration: The authors declare no conflict of interest

Acknowledgements: We would like to thank Miguel Ángel Sebastián, Grégoire Sergeant-Perthuis, Daniel Bennequin, and six anonymous reviewers for crucial feedback on or discussions related to this paper.

Funding/financial support: David Rudrauf was supported by the Swiss National Science Foundation, project funding in Mathematics, Natural sciences and Engineering (division II) #205121_188753

References:


Mach, E. (1897) *Contributions to the analysis of the sensations*. Open Court.


