Some Conceptual and Empirical Shortcomings of IIT[1,2]

Edward F. Kelly

University of Virginia School of Medicine

[Editor's note: The author of this paper first submitted it to the special issue of *Frontiers in Human Neuroscience* that is described in this issue by Weiler et al. The first review he got within a month recommended publication and suggested, but not required, changes. He could not get another review for 6 months, at which point he withdrew it and submitted it to *JAEX*].

Abstract: The Integrated Information Theory of consciousness (IIT) has generated much excitement inside and outside the scientific community, and seems to many the leading contender for a satisfactory theory grounded in systems neuroscience. It is a bold theory, one that provides plausible explanations for various recognized neuroscientific facts, makes surprising predictions that go beyond current scientific orthodoxy but are potentially testable, and has inspired development of what appears to be an effective technique for detecting the presence of consciousness in organisms incapable of verbal report, such as non-human animals, neonates, and severely brain-damaged adults. Despite these virtues, IIT appears fundamentally flawed: This paper first revisits some key conceptual and technical issues that have been raised previously but remain unresolved—in particular, issues concerning IIT's concept of "information" and its approach to the "hard problem"—and then focuses on several empirical phenomena that IIT seems unable to handle satisfactorily. These include: 1. cases of multiple personality or dissociative identity disorder in which complex and overlapping centers of consciousness co-occur in single human organisms; 2. the failure of the intense phenomenology of psychedelic states to be straightforwardly reflected in accompanying neuroelectric activity; and, most critically; 3. the occurrence of profound and personally transformative near-death experiences (NDEs) under extreme physiological conditions such as cardiac arrest, in which IIT predicts that no conscious experience whatsoever should be possible. These empirical arguments show that IIT itself is untenable, and they apply also to its physicalist competitors. Scientifically and philosophically respectable alternatives, however, are available.

Keywords: consciousness, hard problem, reductionism, materialism, realist idealism, MPD/DID, psychedelic neuroimaging, NDEs

Highlights

- Sketches the history and main features of IIT.
- Pinpoints several conceptual weaknesses of IIT.
- Identifies empirical phenomena that falsify IIT and its physicalist competitors.

[1] Address correspondence to: Edward F. Kelly, Ph. D:, Division of Perceptual Studies, Department of Psychiatry and Neurobehavioral Sciences, University of Virginia School of Medicine, 210 10th Street NE, Charlottesville, Virginia 22902, ek8b@virginia.edu

[2] The author thanks Eben Alexander, Etzel Cardeña, Adam Crabtree, Patrick Fowler, Bruce Greyson, Mike Grosso, Bernardo Kastrup, Emily Kelly, Paul Marshall, Michael Nahm, David Presti, Bob Rosenberg, Richard Silberstein, Andreas Sommer, and Max Velmans for helpful comments on earlier versions of this paper.

https://doi.org/10.31156/jaex.24123

- Provides pointers to scientifically and philosophical alternatives to the prevailing physicalist "production" model of consciousness.

The scientific and philosophical study of consciousness was well underway by the early 20th century, in the hands of people such as William James and Henri Bergson, when it was abruptly derailed (and American experimental psychology in general overtaken) by the misguided radical behaviorism of James B. Watson and his followers. Even as late as the 1960s, when I was a graduate student in psychology, the word "consciousness" was rarely spoken in polite scientific company. That aversion has now largely subsided, mercifully, and consciousness research is once again flourishing.

One central aspect of this contemporary renaissance is the emergence of a number of neuroscience-based theories of consciousness, which though differing in many details are generically similar in overall form. Specifically, they all portray the ongoing fluctuations of everyday consciousness as arising from, or at least being closely tied to, dynamically evolving large-scale patterns of neuroelectric activity spanning a widespread core brain system consisting of the thalamus and the interconnected neocortical territories with which it is massively and reciprocally connected.

Integrated Information Theory (henceforth IIT) has become in the eyes of many the leading contender among such theories. This is certainly due in part to its high-profile provenance: Its principal developer, Giulio Tononi, began working on the theory in the 1990s, during his tenure at the Neurosciences Institute in La Jolla with Nobel laureate Gerald Edelman, and he has subsequently been joined in its further development and promotion by Christof Koch, another prominent neuroscientist and the principal protégé of Nobel laureate Francis Crick. Together with various other colleagues, Tononi and Koch have produced numerous papers and books about IIT, and have recently created a non-profit foundation (Tiny Blue Dot) which seems devoted principally to its promotion. IIT has also evidently become the principal driver of the ongoing "adversarial collaboration" among competing neuroscientific theories of consciousness funded by the Templeton World Charities Foundation. Despite its current popularity, however, IIT seems to suffer some very significant liabilities, both conceptual and empirical. To these I now turn, starting with a brief description of the theory itself.

A Very Brief Sketch of IIT

The beginnings of IIT can be found in Edelman & Tononi (2000) and the distillation by Edelman (2003) of its underlying theoretical framework, but the theory has evolved significantly since that early period as Tononi and his later colleagues have progressively elaborated and formalized that original systems-level approach to the neurobiology of consciousness. The historical development can be traced through the accounts provided in Balduzzi & Tononi (2008, 2009), Koch (2012, 2019), Koch et al. (2016), Oizumi et al. (2014), Tononi (2004, 2008, 2012, 2015), Tononi & Koch (2015), Tononi & Laureys (2009), and Tononi et al. (2016). The current version—"IIT 3.0"—was initially set forth by Oizumi et al. (2014) and is described also in Koch (2019) and Tononi (2015). Here I will not describe the theory in detail, but seek only to convey enough of its flavor to provide context for the critical remarks to follow.

IIT begins by attempting to characterize the essential phenomenological properties of conscious human experience (its "axioms"), and then seeks to infer what properties of a physical system such as the brain could potentially support, correspond to, or perhaps "explain" those phenomenological properties (its "postulates"). Following Tononi (2015), IIT's current axioms have been formulated in accordance with the governing principles that they should be about conscious experience itself; self-evident, or immediately given and not requiring derivation or proof; essential to all conscious experience; complete as a set, and thus leaving out nothing vital; consistent, or free of internal contradictions; and independent, in the sense that none can be derived from any combination of the others.

IIT currently recognizes five such axioms. Ever so briefly, intrinsic existence captures the fact that every moment of consciousness is inherently actual or real and experienced from a particular subjective point of view. Composition points to the structure of conscious experiences, which typically incorporate multiple phenomenological properties and distinctions. Information refers to the specificity of each individual experience, which allows it to be differentiated from many others that might have occurred but have not. Integration refers to the unity of each conscious experience, the fact that its potentially many parts are experienced simultaneously as a unified whole. Exclusion means essentially that only one such experience can occur at a time. Note in passing that Edelman & Tononi (2000) had recognized just two principal features of conscious experience, differentiation and integration, and that these remained the key features as late as Koch (2012, pp. 125-126).

IIT holds that for each of its axioms there must be a corresponding property of a physical system such as the brain that can support or account for it. Again, very briefly, intrinsic existence is claimed to arise in connection with any physical system that has causal power over itself (see below, however, for more on this). Composition reflects the fact that the elements of such a system can be co-activated in many possible combinations or subsets. Information or differentiation depends on the size of the repertoire of possible states that a system can assume, and is measured in terms of the degree to which its present state constrains its own past and future states relative to that total repertoire. Integration is effected through causal interactions among the parts of a system, and reflects the irreducibility of its current overall state to interactions among independent subsets of those parts. Exclusion results from the winner-take-all character of system operation, requiring that intrinsic existence be associated only with the sub-state of the current overall state that is maximally irreducible from the perspective of the system itself, neither more nor less.

For more detailed explanations of these complex, difficult, and still-evolving basic concepts so fundamental to IIT, readers should consult the original sources identified above. Their formal development has resulted in three further aspects of IIT that are important for my purposes here. First and most widely known is an information-theoretic quantity termed PHI (Φ, rhymes with "eye"), conceived as the amount of information generated by a system above and beyond that generated by its parts, and understood as a measure of the overall amount of conscious experience associated with the current state of a physical system. Second, and much less well developed, is the notion of a "qualia space," in which the physical substrate of each conscious state exhibits a geometric structure that determines its qualitative character and semantic relations to other such states (Balduzzi & Tononi, 2009). Third is IIT's "Central Identity"—an ontological commitment to identity between a given conscious experience and the maximally irreducible cause-effect structure associated with the system in that state, as embodied in the complex that maximizes Φ (Koch, 2019, pp. 87-89; Oizumi et al., 2014, p. 3).

Development and refinement of the theory has taken place over several decades, using a variety of methods including mathematical formalization of key quantities such as Φ and the study of their behavior in computationally tractable "toy" systems constructed deliberately to exemplify various structural arrangements known to exist in mammalian nervous systems, together with computer simulation studies using increasingly large and realistic model neural networks. One central and very encouraging cumulative result of these efforts has been to show that larger values of Φ occur systematically in conjunction with what are called "small-world" network architectures that combine dense local connections with much sparser long-distance connectivity (Watts & Strogatz, 1998)—precisely the kind of architecture characteristic of mammalian nervous systems (Edelman & Tononi, 2000). The theory thus appears consistent in important respects with known neuroanatomy and neurophysiology. As explained in more detail in the references cited above, it also provides seemingly plausible explanations for various recognized neuroscientific facts. These include, for example, the normal co-occurrence of unified conscious experience with large-scale coordination across cortical domains of brain neuroelectric activity; the relatively minor contribution to consciousness of structures such as the spinal cord, and more importantly the cerebellum with its relatively enormous neural population; the loss of consciousness during general anesthesia, epileptic seizures and deep sleep; and the presence of consciousness during REM sleep.

Surprisingly to me, and controversially, Koch (2012) repudiated his own more conventional earlier views by explicitly rejecting emergence and reductionism in regard to consciousness (p. 119), even going so far as to embrace panpsychism (pp. 131-135). Tononi and Koch (2015) later relaxed that embrace somewhat, perhaps in response to ridicule of it by Searle (2013), but it remains true that IIT, precisely because of its conception of the interiority available to self-influencing physical systems, sees consciousness as far more widely distributed in nature than most scientists have previously imagined. IIT's ultimate philosophical position in regard to the mind/brain relation remains unclear to me, but despite the "Central Identity" noted above it seems (arguably) a constitutive panexperientialism, or perhaps a form of property dualism—positions that recognize irreducible mental properties but view them as depending or supervening on conventional physical properties. Their theory also verges on epiphenomenalism, in that they say very little about what consciousness does, and instead speak almost exclusively about what it's like—essentially a passive accompaniment to the corresponding physical state. The architects of IIT should really be pressed to locate it more precisely within the broader landscape of relevant philosophical possibilities (Marshall, 2021).

Let me immediately acknowledge here, in concluding this brief and inadequate description, that I find a great deal to admire in the work that has gone into the development of IIT. As will become clearer below, I also believe that IIT is correct in taking conscious experience itself as its starting point, in explicitly rejecting emergence and reductionism, and in viewing consciousness as a fundamental feature of nature. At the same time, I also see significant conceptual and empirical problems with the theory in its current form, and turn next to discussion of these.

Conceptual Issues with IIT

The axiomatic foundations of IIT 3.0 have been examined in depth from a philosophical perspective in an important paper by Bayne (2018), who discusses the axiomatic approach in general (which he finds ill-suited to the study of consciousness); the five current IIT axioms individually (which he thinks generally fail to qualify for their intended roles); the unclear logical form of their relationship to the postulates (which is neither deduction nor induction but something more like abduction—read, "guesswork"); and their limited capacity to provide constraints upon a supposedly general theory of consciousness (which results from the near-exclusive focus on human or at least mammalian consciousness). Bayne himself concludes, correctly in my opinion, that the axiomatic foundations of IIT are, in short, "shaky" (p. 7), and Merker et al. (2022), expanding on Bayne's analysis, agree. Here I wish only to discuss a few further points of special interest to me.

One important feature of human conscious experience conspicuously absent from IIT's current axioms is intentionality, the capacity of conscious thoughts and experiences to be "about" or "directed toward" things, events, or states of affairs in the external or internal world. Tononi (2015, footnote 3) explicitly touches upon intentionality as a possible candidate axiom, but immediately dismisses it on grounds that it is (arguably) not a property of all conscious experience. Surely this is too glib. Many ongoing discussions in philosophy of mind revolve around the work of 19th century Austrian philosopher Franz Brentano, who argued that intentionality is *the* mark of the mental, present in all experience, and—more fundamentally—that it is not achievable by any physical system. The jury is still out regarding the universality of intentionality (especially in relation to bare sensory or affective experiences of various kinds), but attempts to naturalize it—that is, to explain it in purely physicalist terms—have generally failed. In sum, whether or not it is present in all experience, intentionality surely must be addressed head-on by any serious theory of consciousness, and IIT has not yet done so. For a lucid and compact introduction to this large literature see Jacob (2019).

Next comes the vexed subject of "information," and the manner in which differing conceptions of information have entered into IIT. Like many other contemporary psychologists and neuroscientists, the developers and expositors of IIT have tended to conflate the technical and everyday meanings of that widely abused term. The technical sense (Shannon & Weaver, 1963) is purely syntactical, and applies strictly to situations in which elements are being drawn with known probabilities from a finite set and sent from a source to a receiver through a noisy channel. The everyday sense, on the other hand, revolves around the meaning or semantic content of conscious experiences and the representational forms such as language and visual imagery with which consciousness routinely operates.

Now, in Edelman & Tononi (2000) the information content of any conscious experience was repeatedly and insistently identified in purely Shannon-like terms with the number of alternative possible experiences ruled out by its occurrence. That identification persists at least as late as Tononi & Laureys (2009, pp. 402-403) and Koch (2012, p. 125), but it is surely incorrect. For one thing, all experiences would be more or less infinitely "informative" in that sense, considering for example that one's world of possible experience might consist just of positive integers, or of real numbers in the interval between zero and one. More importantly, what we normally have in mind in talking about the "information" contained in a conscious experience is surely information in the everyday sense—what that experience itself tells us in positive terms about "what's going on" in our environment.

IIT has by now moved away from Edelman's original view, but not entirely. What we have currently is a kind of hybrid conception: For purposes of calculating an amount of integrated information (Φ), IIT uses a Shannon-like measure of the reduction in uncertainty about the past and future states of a physical system that its current state affords, which can also be understood as a measure of the degree to which the system both affects and is affected by itself. Parenthetically, Bayne (2016) and Pautz (2019) have independently questioned whether it is even meaningful in general to characterize states of consciousness in terms of an overall level or amount of consciousness. Meanwhile, the semantic content or meaning of a conscious state is said to be represented by the geometric structure of its physical substrate in the associated high-dimensional qualia space, in effect translating similarities of meaning into similarities of shape (Balduzzi & Tononi, 2009; Oizumi et al., 2014, supplement; Koch, 2012, pp. 130-131; 2019, pp. 87-89). That latter part of the theory remains relatively undeveloped and exceedingly abstract, and without going into details I have serious doubts about the possibility of developing it in a usefully general form even for relatively simple sensory qualia (see also the second part of Pautz, 2019). Furthermore, even if this could be done, the associated "Central Identity" of IIT—the claim that the semantic content of any conscious experience is equivalent to the geometric shape of the corresponding cause-effect structure in qualia space—seems subject to objections like those of Kripke (1980), who argued essentially that two things cannot be identical if there is anything on one side of the equation not present on the other.

There is something fundamentally unsatisfactory about IIT's approach to the "hard problem," and this lies at the bottom of what troubled philosopher John Searle (2013) in his review of Koch (2012). Searle saw clearly that IIT was attempting to explain consciousness in terms of information in the everyday sense, and argued—correctly, in my opinion—that this is completely circular because information in that sense presupposes consciousness and hence cannot be used to explain it (see also the attempted rebuttal by Koch & Tononi and a further reply from Searle in the March 7, 2013 NYRB). Searle himself left the matter there, but subsequent developments on the IIT side, possibly stimulated by his critique, have made clear exactly how and where consciousness surreptitiously enters the system. Specifically, both Tononi (2015, footnote 7) and Koch (2019, pp. 80-81 and note 3) attempt to justify IIT's crucial first postulate—intended to explain how a physical system can have "intrinsic existence," an associated conscious state organized from its own subjective point of view—by appeal to a Platonic dialogue, Sophist. In that dialogue, the Eleatic Stranger convinces his interlocutor that for anything to be or exist extrinsically—that is, from the point of view of an external observer—all it needs is to have causal power to affect other things. Taking this as their authority, both Koch and Tononi immediately infer that for a physical system to exhibit intrinsic existence of the sort specified by IIT's axiom 1, all it needs is to have causal power over itself—that is, to affect (and be affected by) itself.

That "inference," I submit, is itself pure sophistry. Plato's observer, after all, is already a conscious being, in full possession of intrinsic existence, who is examining something external to himself. The simple fact that a given physical system causally affects itself cannot possibly guarantee interiority of any kind to that system. Coming at this same issue from a different direction, computer scientist Scott Aaronson (2014) demonstrated that simple arrays of logic gates can have non-zero values of Φ, and can be expanded in ways that raise Φ to arbitrarily high levels potentially exceeding those of our own human brains. Whereas Aaronson sees this result as constituting a reductio ad absurdum of the basic premises of their theory, Koch and Tononi take it as a surprising discovery about the distribution of consciousness in nature. Seth et al. (2006) had provided a similar demonstration much earlier, and in Edelman's own lab, but its implications were evidently ignored at the time. Merker et al. (2022) have expanded significantly on Aaronson's theme, but the basic point remains the same.

Additional problems have emerged on the practical side. Calculation of Φ from its formal definition has proved possible only for very small systems, because of a combinatorial explosion in the number of partitions of the system's elements that must be examined. Application to real systems such as brains is further complicated by unclarity about the proper choice of system "elements," which in this case might plausibly be anything from say single neurons to cortical minicolumns or columns, cytoarchitectonic areas, recognized cortical and subcortical "modules" of some sort, or even just the geographic territories underlying a set of EEG electrodes. Barrett & Mediano (2019) have pointed out several further ways in which Φ fails to be well-defined for general physical systems. A number of computationally more tractable approximations have been developed by mathematicians, and further work is underway to explore and refine these, but so far they have tended to behave in mutually inconsistent ways, and at present there is no agreed-upon best way to calculate Φ itself for realistic systems in accordance with its formal definition in IIT (Barrett & Seth, 2011; Mediano et al., 2019).

Meanwhile, however, a separate development associated with the same group but much more loosely tied to the formal theory has proved remarkably successful. I refer here to work on the "Perturbational Complexity Index" (PCI), introduced by Casali et al. (2013) and described more fully by Massimini & Tononi (2018). Their "zap & zip" method is complex in technical detail but straightforward conceptually, using transcranial magnetic stimulation (TMS) to probe the core thalamocortical system directly (bypassing sensory and motor systems and requiring no overt response from the subject), and measuring the brain's consistent response on a subsecond scale using sophisticated multichannel EEG techniques. TMS pulses (the zaps) are applied repeatedly at a particular scalp location, EEG responses averaged and significant cortical sources identified, and the total "complexity" of the cortical response measured using a standard algorithm for data compression (the zip), resulting ultimately in a real number between zero and one. In an international collaboration spanning many years, and including both normal participants in a variety of states (awake, REM sleep, NREM sleep, plus several kinds of anesthesia) and brain-damaged participants of various kinds (stroke victims, plus minimally conscious and locked-in patients), this IIT-inspired research team was able to show in a benchmark sample of 150 cases that a threshold PCI value of .31 discriminated perfectly between conscious and unconscious states—that is, with 100% sensitivity (no false negatives) and 100% specificity (no false positives). Note that this sample included some persons who were unable to report being awake at the time of testing but could do so later, such as locked-in patients who later recovered and normal individuals exposed to subanesthetic doses of ketamine.

2022, Vol. 2, No. 2, pp. 223-263

These initial results with the PCI may not hold up perfectly in future work, inasmuch as a dozen or so of the 150 benchmark cases lie perilously close to the magical .31 boundary (Massimini & Tononi, 2018, p. 124), but they are already of potentially great practical significance, and further work is underway to optimize various parameters of the technique such as the locus, magnitude and rate of stimulation and details of response measurement. What is most significant for my purposes here, however, is the loose connection to IIT proper: The PCI does not rely at all on the formal definition of Φ, but instead more directly captures what Edelman & Tononi (2000) had originally identified as the two most critical properties of large-scale patterns of neuroelectric activity in their "dynamic core" (the thalamocortical system)—differentiation and integration. In essence, PCI grows large when TMS-driven neural activity spreads in differentiated form to remote locations, but assumes low values either if the response remains localized, or if it spreads widely but only in undifferentiated or stereotypical form.

This saga of Φ versus the PCI exemplifies something that for me has come to seem characteristic of IIT in general as a theory—specifically, that what is really new (the elaborate formalism) is not all that useful or good, while what is really good (directly measuring relevant complexities in brain activity) is not all that new. I turn next to additional challenges to IIT of empirical sorts.

Empirical Issues

Disconnection Syndromes and the IIT Exclusion Postulate

The IIT exclusion postulate essentially asserts that a physical system such as the brain can normally accommodate just one conscious state at a time—specifically, the conscious state associated with whatever sub-state of that system maximizes Φ. Note that this picture is consistent with the prevailing modern view in psychology and neuroscience that everyday consciousness is normally all the consciousness there is, supported by various fast, massively parallel, and unconscious forms of brain activity. What does IIT predict, then, if parts of the underlying physical system cease to interact in normal ways?

There are two main types of cases to consider. The first involves anatomical disconnection, in particular the disconnection of cerebral hemispheres effected by drastic "split-brain" surgeries in epileptic patients, which are carried out rarely and as a last resort to prevent interhemispheric spread of seizures in otherwise intractable cases. In this situation, IIT clearly predicts that the result will be two distinct conscious minds or personalities inhabiting the same skull, and developers and proponents of IIT from Edelman & Tononi (2000) onward, relying primarily on the early work of Roger Sperry and Michael Gazzaniga, have generally acted as though this is an established fact. In reality, however, the results of such surgeries are much more complicated and uncertain. The early work certainly established that the left and right hemispheres have specialized abilities that can only be revealed by careful behavioral testing. Things also definitely happen that seem potentially consistent with the two-minds picture, such as the patient blushing when the nonverbal right hemisphere is shown an embarrassing photo that the verbal left hemisphere denies having seen, or when the left hand tries to interfere with something the right hand is doing. But the overwhelming impression of most observers—including, importantly, the patients themselves—is that practically nothing has changed. It was recognized early on that appearances of persisting unity might simply reflect incomplete separation of the hemispheres through sparing of structures such as the anterior and posterior commissures, but this cannot account for all cases, and substantial scholarly arguments for continued unity have been advanced for example by John Eccles (Popper & Eccles, 1977, chapter E5), Pinto et al. (2017), and Nagel (1979). Even Michael Gazzaniga himself has more recently expressed uncertainty about the actual state of things, particularly in regard to the mute right hemisphere (Gazzaniga & Miller, 2009). In sum, the existing split-brain cases are so complicated that we do not really know at this point how best to interpret them (de Haan et al., 2020). See also Nahm et al. (2017) for some cases of related type such as hemispherectomies. Koch (2019, pp. 108-111) also brings up the interesting subject of unusual connection syndromes, as for example between the brains of craniopagus conjoined twins. Such cases are if anything rarer and more complicated than the split-brain cases, but some may be particularly difficult for IIT: The long-lived Schappell twins, for example, are conjoined at the frontal lobes yet have distinct streams of consciousness and drastically divergent minds and personalities.

Cases of the second type, involving functional disconnection, are both more common and more clear-cut in their implications. I refer here in particular to the large literature concerning "multiple personality" or "dissociative identity" disorders (MPD/DID), in which two or more distinct personalities or streams of consciousness appear to be associated with a single biological organism. IIT recognizes the possibility of such cases, but places severe restrictions on their possible form due to its exclusion postulate. In particular, the physical substrate of any stream of consciousness operating concurrently with the main or dominant stream—the everyday conscious self—cannot overlap with that of the dominant stream. Furthermore, it will in general be diminished in complexity relative to that main complex, and hence capable only of supporting a correspondingly reduced secondary consciousness (Koch, 2019, p. 112).

These predictions of IIT, and that mainstream modern picture of the psyche more generally, are challenged by facts that had already come to light during the rich but now largely forgotten early history of dynamic psychiatry, with its demonstrations that "unconscious cerebration" could not fully account for the properties of psychological automatisms and secondary personalities (Crabtree, 2007; Ellenberger, 1970). Secondary centers of consciousness often appear to be full-scale minds or personalities at least on par with the normal waking self. They also sometimes clearly operate concurrently with that self, rather than simply alternating with it, and share with it fundamental cognitive capacities such as linguistic fluency, which implies that the associated neural substrates almost certainly overlap. They are also sometimes conspicuously more complex and able than the primary personality, as for example in the extraordinary case of Patience Worth (Prince, 1964). In addition, and worst of all from the point of view of IIT, "alter" personalities A and B sometimes display an asymmetric form of co-consciousness such that B is aware of much that goes on in A's conscious experience at the time it is happening, but not vice-versa (Braude, 1991; Janet, 1889; Prince, 1908).

British psychical researcher F. W. H. Myers (1903) drew upon these and other unusual human psychophysical capacities to develop an expanded picture of human personality in which the everyday consciousness is included within a normally hidden consciousness of greater capacities and wider scope to which it typically has little or no direct access. Analytical philosopher Stephen Braude (1991) arrived independently at essentially the same picture, and specifically in relation to cases of multiple personality. It has not yet been adequately appreciated that William James deliberately and explicitly adopted Myers's model for his own explanatory purposes in The Varieties of Religious Experience, and that a century-plus of subsequent research has strengthened the evidence supporting it on multiple fronts including MPD/DID as well as extreme psychophysical influence, some key properties of human memory, near-death and related experiences, genius, and mystical experience, as shown in Kelly et al. (2007). This more inclusive subliminal consciousness is precisely the "entirely unsuspected peculiarity in the constitution of human nature" to which James alludes in Chapter 10 of VRE, drawing upon the work of Myers, Janet and Binet. In an address to the Society for Psychical Research, of which he was a member, Sigmund Freud later characterized Myers's construct—incorrectly—as amounting to an "unconscious conscious." See Crabtree (2007, pp. 327-332) for a detailed account. As will soon become evident, this expanded psychological model has deep connections with our remaining topics as well. Much more is at stake here than just IIT's exclusion postulate.

Psychedelic Neuroimaging

The recent renaissance of research on psychedelics (Pollan, 2018), coupled with decades of advances in functional neuroimaging technology, has opened a path toward well-controlled experimental investigation of mystical states of consciousness and their physiological accompaniments. Surely, for persons like Myers and James, this amounts to a scientific dream come true, the possibility of studying brain-mind correlation across an unusually wide range of conscious states. The subject is still in its infancy, but surprising and challenging results have already come to light.

The natural if perhaps naïve expectation is that the intense phenomenology of psychedelic experiences will likely be accompanied by some sort of parallel intensification readily observable in accompanying patterns of neural activity. This expectation seemed to be confirmed early on by experiments using oral doses of psilocybin in conjunction with positron emission tomographic (PET) neuroimaging (see Kelly et al., 2007, pp. 542-553). But then came the study by Carhart-Harris et al. (2012), who used venous injection of psilocybin in combination with two kinds of functional magnetic resonance imaging (ASL and BOLD fMRI, for measurement of blood flow and blood oxygenation, respectively), thereby enabling efficient tracking of a shorter-lasting but intense psychedelic state using neuroimaging techniques having far better temporal and spatial resolution than PET. Most observers including Christof Koch were shocked by the results (see https://www.scientificamerican.com/article/this-is-your-brain-on-drugs). No increases in activation were observed anywhere in the brain. Instead, prominent decreases were observed, most strikingly in a brain system called the Default Mode Network (DMN), which is known to be especially active in the resting state and which in effect represents the neural embodiment of the Freudian ego (Carhart-Harris & Friston, 2010). Major nodes or hubs of the DMN such as anterior and posterior cingulate cortex (ACC and PCC), medial prefrontal cortex (mPFC) and thalamus were strongly deactivated, with average reductions on the order of 10-15%, and the reported intensity of the experience correlated positively and strongly with the magnitude of those reductions. The DMN also appeared to lose functional integrity, with mPFC and PCC in particular becoming decoupled. A subsequent study from the same group using the same drug protocol in conjunction with magnetoencephalographic (MEG) neuroimaging revealed sharp decreases in oscillatory power across a wide range of frequencies in the same cortical regions (Muthukumaraswamy et al., 2013), and similar effects have more recently been found in a number of further studies by the same and other research groups, and using additional agents such as DMT/Ayahuasca, LSD, and ketamine (Palhano-Fontes et al., 2015; Carhart-Harris et al., 2016). These unexpected effects are large and robust, and I will return to them shortly.

Perhaps driven by their own theoretical expectations, the Carhart-Harris group in particular has continued to search, without much success, for something in the brain's neuroelectric activity that consistently increases in conjunction with the "higher" states of consciousness induced by psychoactive agents. That effort culminated, at least temporarily, in the publication of Schartner et al. (2017), the abstract of which states without qualification that they have found "reliably higher spontaneous signal diversity" in MEG signals accompanying states of consciousness induced by psilocybin, ketamine and LSD. That claim, however, is not warranted by the reported results, which derive from reanalysis of MEG data collected in previous experiments by the same group. Their measure of "signal diversity" is similar to the previously discussed Perturbational Complexity Index (PCI; Casali et al., 2013) in that it varies between 0.0 and 1.0 and reflects the complexity or incompressibility of the estimated cortical sources of surface-recorded data, but their technique is applied to individual 2-second segments of spontaneous MEG rather than to time-averages of responses to repeated TMS pulses (becoming, in effect, "zip-without-zap"). The key results appear in their Fig. 2. Mean diversities for the ordinary waking (placebo) conditions are all around .98 or slightly above. The mean diversities for the drug conditions are indeed all higher, but the differences between drug and placebo means are tiny in absolute magnitude—on the order of .005 or less—and not even statistically significant for psilocybin. For some participants in all drug conditions the mean differences from placebo are essentially zero, or sometimes even in the wrong direction, and the underlying raw diversity distributions all strongly overlap. The high significance levels reported for some statistical tests of the tiny differences between drug and placebo mean diversities were achieved only because of the extremely large Ns available for the tests—many hundreds to thousands, apparently, although they are not specifically stated in the report—and this depends upon the dubious use of serially correlated MEG segments rather than participants as the unit of analysis. In sum, signal "diversity" as these authors define it is hardly the robust discriminator advertised in the paper's abstract.

The fact that Schartner et al. (2017) did not succeed in identifying a genuinely robust discriminator does not entail that no such discriminator exists. It bears emphasis here that this leading-edge group, like many others, is working at the rapidly advancing frontier of functional neuroimaging technology, where everybody now understands clearly that the key to the whole subject lies in finding improved ways of characterizing and tracking large-scale neural dynamics on a time-scale relevant to experience and behavior. Numerous other EEG/MEG methods for doing that are either already available or under development somewhere, and perhaps one or more of those other methods can do a better job of detecting relevant changes in the brain activity induced by psychoactive substances. The PCI itself is certainly a good prospect, and was recommended also by Gallimore (2015), who used IIT to predict that psychedelics will result in increased diversity or differentiation, but argued that integration needs to be brought into the measurement system as well (which of course the PCI implicitly does). Another good reason is that the maximum value of the PCI in ordinary waking states is only around .7 (Massimini & Tononi, 2018, p. 124), affording far more room for it to grow in "higher" states. Other candidates might include the many existing measures of multichannel signal "entropy"; the slope of the 1/f portion of the EEG power spectrum (thought to reflect the local balance of inhibition and excitation; He, 2014); "causal density" (average values of directed connectivity; Seth et al., 2008); global descriptors of multichannel signal amplitude, frequency and complexity (Wackermann & Allefeld, 2009); and others. The brute fact of the matter is that we do not yet have a very solid collective grip on optimal measures of neural dynamics and their relations to conscious experience. It is also true, given the well-known non-linearity of brain dynamics, that unimpressive-looking changes in measures of brain activity could potentially result in large functional effects. Nevertheless, for me as a long-time EEG researcher it remains profoundly surprising and puzzling, given the ease with which EEG discriminates between relatively mundane conditions such as waking and drowsy states, that it has so far proven so difficult to find anything in surface-recorded neuroelectric activity that distinguishes clearly and reliably between ordinary wakefulness and the extraordinarily intense and sometimes life-transforming states of consciousness induced by these psychoactive substances.

Meanwhile, the most robust finding to date is the relatively dramatic deactivation and fragmentation of the DMN produced by psychedelics (Seth et al. (2018) attempted to distance themselves from this finding, but co-author and group leader Carhart-Harris (2018, p. 174) subsequently reaffirmed it). Parenthetically, similar DMN-suppression effects have been found in advanced meditators by Brewer et al. (2011), and much additional evidence along the same lines can be found in Goleman & Davidson (2017, chapter 8).

What are we to make of all this? The full historical context for this question is spelled out in a valuable paper by Swanson (2018), who goes back as far as the 19th century and specifically includes early figures such as Myers, James and Bergson under the heading of "filtration" theorists, who interpret the profound psychological effects of psychedelics as incursions from deeper and normally hidden parts of the mind as a result of disruption of some sort of "filtering" or "reducing valve" mechanism that normally confines conscious mentation within limits adapted to the needs of everyday life. Such theorists have differed widely in terms of their conceptions of the nature of those normally hidden parts of the mind, with contemporary filter-type theorizing of course almost invariably seeking to understand everything in purely reductionist, brain-based terms. Drawing upon his own background in psychoanalytic thought, for example, Carhart-Harris has argued that the old but still-useful Freudian distinction between secondary and primary processes maps onto the distinction between ego-like functions exercised by the DMN and the activity of more primitive brain areas that it normally controls, and that by disrupting the DMN and thus "dissolving" the ego, psychedelics release sources of primary-process material distributed widely across other parts of the brain (Carhart-Harris & Friston, 2010; Carhart-Harris et al., 2014). This is their neuroscience-based version of a filter-type theory, which they believe is capable of accounting for all relevant mental phenomena including both pathological (such as pre-psychotic mentation) and supernormal (such as high forms of creative thinking). Carhart-Harris & Friston (2010) even congratulate themselves for "addressing topics which have hitherto been considered incompatible with the cognitive paradigm" (p. 1275).

Swanson (2018) adopts the same reductive point of view, presenting Carhart-Harris's "entropic brain hypothesis" (EBH) together with IIT and hierarchical predictive processing (HPP) as the principal current examples of filtration theories couched in contemporary cognitive neuroscientific terms (parenthetically, Carhart-Harris & Friston (2019) have more recently attempted to integrate EBH and HPP under the umbrella of a unified REBUS or "relaxed beliefs under psychedelics" model of psychedelic interactions, leaving IIT aside). All these theories presume, axiomatically, that everything that enters consciousness in richer-than-normal psychedelic or mystical states of consciousness must come from elsewhere in the brain, either directly (from whatever is already stored there) or indirectly (by way of our sensory surfaces). The earlier filtration theorists, however—including Myers, James, and Bergson—all felt compelled by evidence to adopt a more radical view. Consciousness, they argued, "overflows the organism" (Bergson) and is ultimately grounded in some sort of transpersonal "Mind at Large" (Aldous Huxley). Swanson (2018) declines to discuss this ontologically more extreme version of filter theory, but Kelly & Presti (2015) have specifically embraced it in the context of psychedelic experiences and mystical experiences more generally, pointing out among other things that it can potentially explain the large amount of historical and cross-cultural evidence linking such states with unusually strong outbreaks of paranormal or "psi" phenomena. The term "psi" refers here in theoretically neutral fashion to the various categories of paranormal effects—"extrasensory perception" or "ESP" on the input side (including telepathy, clairvoyance, precognition and retrocognition), and psychokinesis or "mind over matter" effects on the output side. The defining property in all cases is that information appears to flow between an organism and its environment despite the presence of barriers (such as physical shielding, or distance in space and/or time) that would be expected on conventional physical principles to prevent such flows from occurring. See also Cardeña, 2014, 2018, 2020; Cardeña & Winkelman, 2011; Cardeña et al., 2015, chapter 12; Kelly et al., 2007, pp. 522-523; Kelly & Locke, 2009; Marshall, 2011). The need for serious consideration of such an expanded picture of the human psyche becomes inescapable in connection with the final empirical phenomenon to be discussed.

Near-Death Experiences (NDEs) Occurring under Extreme Physiological Conditions

Decades of experimental and theoretical work in neuroscience and psychology have led to a very strong consensus among contemporary mainstream neuroscientists

2022, Vol. 2, No. 2, pp. 223-263

that all the varying states of human consciousness are associated somehow with correspondingly varying patterns of neuroelectric activity in the brain. Different groups have seized upon different possible "markers" of consciousness, such as gamma waves, late components of evoked potentials, desynchronization or "activation" of ongoing EEG, and the PCI (Koch et al., 2016), but all agree that consciousness can only occur in brains that are currently capable of generating oscillatory activity in the frequency range of say 8 to 70 Hz or more and flexibly coordinating that activity over large distances in neocortex. Different theorists hold somewhat differing views as to the precise role such activity might play in generating, shaping, integrating and sustaining consciousness, but all agree that its availability is a necessary condition for conscious experience of any sort to occur. IIT itself is representative of this broader consensus.

However, it is now well established that so-called near-death experiences (NDEs) sometimes occur under extreme physiological circumstances such as deep general anesthesia and/or cardiac arrest in which those supposedly necessary conditions for consciousness have been severely degraded or abolished altogether. A primary purpose of general anesthesia, after all, is to render surgical patients unconscious of their surgeries, and the PCI and various other measures confirm that this is what normally happens; adequately anesthetized persons enter physiologically distinctive states different from those that accompany normal waking consciousness, and have nothing to report following their recovery from anesthesia. In the more extreme case of cardiac arrest, blood flow to the brain drops almost instantaneously to zero, the EEG flat-lines within 10 to 20 seconds, and even neuronal action-potentials—the ultimate physical basis for the causal interactions among brain elements upon which IIT rests—are quickly suppressed. This guarantees that in cardiac arrest the cortex is not merely inactive but deactivated, with the result that Φ necessarily goes to zero. Nevertheless, something on the order of 10-20% of such patients report having had not only a conscious experience of some sort, but the most profound and potentially transformative experience of their entire lives (Greyson, 2021 a, b; Holden et al., 2009; Kelly et al., 2007, chapter 6; van Lommel et al., 2001).

2022, Vol. 2, No. 2, pp. 223-263

For scientists determined to resist the theoretical implications of such observations, the first line of defense is typically an objection to the effect that even in the presence of a flat-lined EEG there might be some sort of residual neuroelectric activity going on, not visible in scalp EEG, that could explain the conscious experiences. That objection, however, completely misses the mark: The issue is not whether there is brain activity of any kind whatever, but whether there is brain activity of the specific form regarded by contemporary neuroscience as a necessary condition for conscious experience, and activity of that form is readily detectable in scalp EEG (parenthetically, existing attempts to explain NDEs in terms of "brain flashes"—supposed surges of neuroelectric activity at or near the point of death—are of little scientific value; see Greyson, 2012b, pp. 31-32).

The next line of defense is to propose that the reported experiences did not actually occur during the period of apparent unconsciousness, when the brain was severely impaired, but either before or after the impairment. This objection, however, seems to be ruled out in many cases by the presence of time anchors, aspects of the reported experiences that can be verified as having happened during the period of apparent unconsciousness, but that could not have been observed by the patient even if fully conscious in the normal way. These include, for example, things reported as being seen or heard despite deliberate pre-surgical blockage of the corresponding senses, or events occurring during that time period but in a physically remote location. Over a hundred such cases are reported in Holden et al. (2009, pp. 185-211), and many more in Rivas et al. (2016). Many are well documented, and they cannot be collectively dismissed as mere "anecdotes." It would be desirable, of course, to supplement these case collections with experimental evidence based for example on perception by patients of targets deliberately set out in cardiac surgery theaters. There have been a few preliminary attempts along these lines, mostly unsuccessful (see Parnia et al., 2014), but we need more creative solutions to the problems of getting targets displayed in places where clinicians do not want them, and getting patients to seek them out and remember them in the midst of a life-threatening crisis.

Time anchors typically involve psi-type occurrences, and determined skeptics will try to ignore these, regardless of the level of documentation involved, on grounds that they are insufficient on their own to guarantee the reality of psi. That objection is irresponsible, however. The reality of psi as a fact of nature has been independently established by the large amounts of relevant observational and experimental evidence that have gradually accumulated, starting even before the founding of the British Society for Psychical Research in 1882. What is unusual about these NDE cases concerns only the circumstances in which psi events are occurring.

One final escape route, available only to persons who accept the reality of psi, is to agree that the reported experiences are imaginative constructions dating from a time before or after the period of brain impairment, but to admit in addition the capacity of psi processes to explain the time anchors—that is, via precognition from a time before, or retrocognition from a time after, the impairment. Even this interpretation, however, is contradicted by the fact that the reported experiences always precede, and never follow, the reports themselves.

For additional pointers into the very large body of evidence for psi phenomena in general see for example Cardeña (2018), Radin (2006) and the annotated bibliography of Kelly et al. (2007). For extensive further discussion of the "real-time" vs. "reconstruction" models of NDEs see in particular Nahm & Weibel (2020), who emphasize that out-of-body experiences (OBEs), including perceptions of the percipient's own body from a different spatial location (autoscopy), occur commonly not only in NDEs but under other conditions as well, and typically occur unambiguously in real time. They also underscore the additional challenge posed to physicalist accounts of both NDEs and OBEs by the fact that highly similar experiences occur under such an extraordinary variety of physiological circumstances.

For fuller documentation and justification of my brief treatment here of this theoretically crucial subclass of NDEs, readers are encouraged to consult the much richer sources cited above. These cases come straight from the heartland of contemporary biomedical science, and more and better such cases are coming to light as advances in resuscitation medicine enhance our ability to retrieve dying persons from the borderlands of death. Their special significance lies in demonstrating that unusually intense conscious experience sometimes occurs in association with brain conditions that IIT along with every other existing neuroscientific theory of consciousness deems incapable of supporting it. But as I will next briefly explain, the ontologically more radical filtration theory of Myers, James and Bergson allows us to come to grips with these NDE cases and many other "rogue" empirical phenomena in a scientifically and philosophically respectable way.

Conclusion

William James (1900), with characteristic precision, put his finger on the essential logical point. The overall correlation generally observed between mental happenings and happenings in the brain—which everybody accepts—can be interpreted in more than one way. Physicalists see it as straightforwardly confirming their "production" model, according to which mind and consciousness are manufactured by neurophysiological processes occurring in brains, in something like the way a tea kettle generates steam, or the electric current flowing in a lamp generates light. But the true function of the brain might instead be permissive, like the trigger of a crossbow, or transmissive, like an optical lens

2022, Vol. 2, No. 2, pp. 223-263

or prism, or like keys of a pipe organ—or perhaps in more contemporary terms like the receivers in our radios and televisions. The observed correlations of mind and brain would then reflect the operation of some sort of mental reality (which in James's view could be anything from a finite mind or personality to some sort of World Soul) that normally is closely coupled to the brain functionally, but somehow distinct from the brain and potentially capable of operating on its own. Within this basic framework James himself spoke variously of the brain as straining, sifting, canalizing, limiting, and individualizing that larger mental reality existing behind the scenes, and portrayed the brain as exerting these various effects in a manner dependent on its own functional status. James argued further that this picture is in principle compatible with all the facts conventionally interpreted under the production model, and that however metaphorical and incomprehensible it might at first seem, it is in reality no more so than its materialist rival. It also has certain positive superiorities, in particular its potential to explain various additional facts then being unearthed by F. W. H. Myers and his colleagues in psychical research—and numerous related facts unearthed since that time, as indicated above (Kelly et al., 2007).

Among the empirical phenomena highlighted in the present paper, NDEs occurring under extreme physiological conditions including cardiac arrest are surely the most critical, theoretically. Significant advances in science are standardly understood to derive from confirmation of bold conjectures and falsification of cautious or timid ones (Chalmers,1994, pp. 54-55), and NDEs of this sort exert both types of effects at once in favor of the Myers/James picture. On the one hand, conventional physicalist models including IIT clearly predict that no experience whatsoever should be possible under these conditions, and this "cautious" prediction is falsified by the experiences that do occur. From the Myers/James filter perspective, meanwhile, the possibility of such experiences can certainly be anticipated, if not strictly "predicted" as if by some sort of deductive logic, and this "bold" conjecture is confirmed. Facts of this theoretically pivotal sort cannot go on being ignored by mainstream science; we cannot and will not arrive at a satisfactory theory of consciousness unless we begin with a synoptic empiricism that embraces all theoretically relevant facts.

The existing empirical challenges to the currently prevailing physicalism are profound and inescapable, and they justify thoroughgoing re-evaluation of that physicalist worldview itself. The Myers/James picture as described so far is basically a psychological theory, but it has definite consequences for our picture of reality more generally. It is essential to appreciate here, first, that physicalism as conventionally understood is not itself science but metaphysics—a philosophical position deriving from, and chained to, the classic Newtonian physics of the late 19th century. That seemingly secure foundation in physics has subsequently eroded, however, beginning with the rise of relativity and quantum theories, with the result that physicalism is no longer consistent with our deepest physical science. Indeed, "matter" itself as classically conceived does not exist, and space and time can no longer be conceptualized as providing a pre-existing container for events. Furthermore, we have no credible understanding of how consciousness could be manufactured by physical processes occurring in brains, and recent work in philosophy of mind has convinced many that we can never achieve one (Chalmers, 1996; Nagel, 2012).

The underlying basic question here is not whether we will have metaphysics—because we inevitably will, whether conscious of it or not—but whether we will have good metaphysics or bad. And the second key point to appreciate here is that scientifically and philosophically respectable alternatives to physicalism are already on offer. I believe the great pioneers of consciousness research including Myers, James and Bergson were on the right track with their ontologically more radical form of "filtration" theory, and that the arguments and data presented in this paper (along with the other sources cited) drive us beyond IIT and its physicalist competitors in the direction of just such a theory. Descriptions and evaluations of various conceptual frameworks or worldviews or metaphysical systems of this sort both ancient and modern that seem capable of accommodating phenomena resistant to physicalist explanation can be found in Eastman (2020), Goff (2017), Kastrup (2019), Kelly et al. (2007, 2015), Kelly & Marshall (2021), Marshall (2021), Moreira-Almeida & Santos, (2013), Seager (2020), Strawson (2006), and numerous other sources pointed to therein. Without going into detail here, the central tendency of these theoretical possibilities, as anticipated by James in The Varieties of Religious Experience and his own later philosophical work, is toward a form of pluralist idealism falling generally within the category of evolutionary panentheisms (Hartshorne & Reese, 2000; Kelly et al., 2015). These are all conceptual near-neighbors of IIT, ironically, in making consciousness more than a product of our individual brains and much more fundamental to the scheme of reality as a whole, but arrived at from radically different directions. Many are also demonstrably more consistent with modern physics than physicalism itself.

For an account of James's trajectory see Kelly et al. (2015, Chapter 14). My personal sympathies presently tend toward some form of priority monism (Schaffer, 2010), possibly relying in particular upon an emerging conception according to which both classical and quantum physics can be derived from more basic "informational" principles which themselves are grounded in an ontologically prior universal consciousness (D'Ariano & Faggin, 2021; Faggin, 2021). Note that this program, if successful, will in effect drain physics itself of its traditional "physical" content, and move it strongly in the direction of realist idealism. Another live alternative is the various forms of dual-aspect monism, which seek to capitalize upon the ability of conventional physics to explain many regularities of human experience while avoiding a possible "inverse hard problem" of deriving matter from consciousness (e.g., Marshall, 2021; Velmans, 2021).

Let me underscore in conclusion that my theoretically revisionist colleagues and I genuinely celebrate existing science, and are only attempting to expand it in ways that will permit it to accommodate a wider range of humanly important empirical phenomena. What we are advocating is not any form of supernaturalism but an expanded naturalism. For an especially well-informed and optimistic account of the positive implications of these emerging views for the science of consciousness, including numerous suggestions for further basic and applied research, see Presti (2021).

The bottom line? IIT is bold, yes, but it is not bold enough!

References

Aaronson, S. (2014, May 30). *Why I am not an integrated information theorist (or, the unconscious expander).* www.scottaaronson.com/blog

Balduzzi, D., & Tononi, G. (2008). Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLOS Computational Biology, 4* (6): e1000091. https://doi.org/10.1371/journal.pcbi.1000091

Balduzzi, D., & Tononi, G. (2009). Qualia: The geometry of integrated information. *PLOS Computational Biology, 5*(8): e1000462. https:/doi.org/10:1371/journal.pcbi1000462

Barrett, A. B., & Mediano, P. A. M. (2019). The Phi measure of integrated information is not well-defined for general physical systems. *Journal of Consciousness Studies, 26*(1-2), 11-20.

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLOS Computational Biology, 7* (1). https://doi.org/10.1371/journal.pcbi.1001052

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. Neuroscience of Consciousness, 4(1), 1-18. https://doi.org/10.1093/nc/niy007

Bayne T., Hohwhy, J., & Owen, A. M. (2016). Are there levels of consciousness*? Trends in Cognitive Sciences, 20*(6), 405-413. https://doi.org/10.1016/j.tics.2016.03.009

Braude, S. E. (1991). *First person plural: Multiple personality and the philosophy of mind.* Routledge.

Brewer, J. A., Worhunsky, P. D., Gray, J. R., Tang, Y-Y., Weber, J., & Kober, H. (2011). Meditation experience is associated with differences in default mode network activity and connectivity. *Proceedings of the National Academy of Sciences USA, 108*(50), 20254-20259. https:doi.org/10.1073/pnas.1112029108

Cardeña, E. (2014). A call for an open, informed study of all aspects of consciousness. *Frontiers in Human Neuroscience, 8*(17). https:doi.org/10.3389/fn-hum.2014.00017

Cardeña, E. (2018). The experimental evidence for parapsychological phenomena: A review. *American Psychologist, 73*(5), 663-677. https://doi.org/10.1037/amp0000236

Cardeña, E. (2020). Derangement of the senses or alternate epistemological pathways? Altered consciousness and enhanced functioning. *Psychology of Consciousness: Theory, Research, and Practice, 7*(3), 242-261. https://doi.org/10.1037/cns0000175

Cardeña, E., Palmer, J., & Marcusson-Clavertz, D. (Eds.) (2015). *Parapsychology: A handbook for the 21st century*. McFarland.

Cardeña, E., & Winkelman, M. (Eds.) (2011). *Altering consciousness: Multidisciplinary perspectives* (2 vols.). Praeger/ABT-CLIO.

Carhart-Harris, R. L., Erritzoe, D., Williams, T., Stone, J. M., Reed, L. J., Colasanti, A., & Nutt, D. J. (2012). Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences USA, 109*(6), 2138-2143. https://doi.org/10.1073.pnas.1119598109

Carhart-Harris, R. L., & Friston, K. J. (2010). The default-mode, ego-functions and free-energy: A neurobiological account of Freudian ideas. *Brain, 133*(4), 1265-1283. https://doi.org/10.1093/brain/awq010

Carhart-Harris, R. L., & Friston, K. J. (2019). REBUS and the anarchic brain: Toward a unified model of the brain action of psychedelics. *Pharmacological Reviews, 71*(3), 316-344 https://doi.org/10.1124/pr.118.017160

Carhart-Harris, R. L., Leech, R., Hellyer, P. J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D. R., & Nutt, D. J. (2014). The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience, 8*(2), 1-22. https://doi.org/10.3389/fnhum.2014.00020

Carhart-Harris, R. L., Muthukumaraswamy, S., Roseman, L., Kaelen, M., Droog, W., Murphy, K., & Nutt, D. J. (2016). Neural correlates of the LSD experience revealed by multimodal neuroimaging. *Proceedings of the National Academy of Sciences USA, 113*(17), 4853-4858. https://doi.org/10.1073/pnas.1518377113

Carhart-Harris, R. L. (2018). The entropic brain – revisited. *Neuropharmacology, 142*(11), 167-178. https://doi.org/10.1016/j.neuropharm.2018.03.010

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine, 5*(198), 1-11. https://doi.org/10.1126/scitranslmed.3006294

Chalmers, A. F. (1994). *What is this thing called science?* (2nd ed.). Hackett. (Original work published 1976)

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Crabtree, A. (2007). Automatism and secondary centers of consciousness. In E. F. Kelly, E. W. Kelly, A. Crabtree, A. Gauld, M. Grosso, & B. Greyson (Eds.), *Irreducible mind* (pp. 301-365). Rowman & Littlefield.

D'Ariano, G. M., & Faggin, F. (2021). *Hard problem and free will: An information-theoretical approach.* arxiv:2012.06580v2. https://doi.org/10.1007/978-3-030-85480-5_5

De Haan, E. H. F., Corballis, P. M., Hillyard, S. A., Marzi, C. A., Seth, A., Lamme, V. A. F., ... & Pinto, Y. (2020). Split-brain: What we know now and why this is important for understanding consciousness. *Neuropsychology Review, 30*, 224-233. https://doi.org/10.1007/s11065-020-09439-3

Eastman, T. E. (2020). *Untying the Gordian knot: Process, reality, and context.* Rowman & Littlefield.

Edelman, G. M., & Tononi, G. (2000). *A universe of consciousness: How matter becomes imagination.* Basic Books.

Edelman, G. M. (2003). Naturalizing consciousness: A theoretical framework. *Proceedings of the National Academy of Sciences, 100*(9), 5520-5524. https://doi.org/10.1073/pnas.093149100

Ellenberger, H. F. (1970). *The discovery of the unconscious: The history and evolution of dynamic psychiatry*. Basic Books.

Faggin, F. (2021). Consciousness comes first. In E. F. Kelly & P. Marshall (Eds.), *Consciousness unbound* (pp. 283-319). Rowman & Littlefield.

Gallimore, A. R. (2015). Restructuring consciousness - the psychedelic state in light of Integrated Information Theory. *Frontiers in Human Neuroscience, 9,* 1-16. https://doi.org/10.3389/fn-hum.2015.00346

Gazzaniga, M. S., & Miller, M. B. (2009). The left hemisphere does not miss the right hemisphere. In S. Laureys & G. Tononi (Eds.), *The neurology of consciousness* (pp. 201-210). Academic Press. https://doi.org/10.1016/B978-0-12-374168-4.00020-4

Goff, P. (2017). *Consciousness and fundamental reality.* Oxford University Press. https://doi.org/10.1093/oso/9780190677015.001.0001

Goleman, D., & Davidson, R. J. (2017). *Altered traits: Science reveals how meditation changes your mind, brain, and body*. Avery Publishing.

Greyson, B. (2021a). *After: A doctor explores what near-death experiences reveal about life and beyond.* St. Martin's Essentials.

Greyson, B. (2021b). Near-death experiences. In E. F. Kelly & P. Marshall (Eds.), *Consciousness unbound* (pp. 17-56). Rowman & Littlefield.

Hartshorne, C., & Reese, W. L. (Eds.). (2000). *Philosophers speak of GOD.* Humanity Books. (Original work published 1953)

He, B. J. (2104). Scale-free brain activity: past, present, and future. *Trends in Cognitive Sciences, 18*(9), 480-487. https://doi.org/10.1016/j.tics.2014.04.003

Holden, J. M., Greyson, B., & James, D. (Eds.) (2009). *The handbook of near-death experiences.* Praeger/ABC-CLIO.

Jacob, P. (2019). Intentionality. In E. N. Zalta (Ed.), *Stanford encyclopedia of philosophy* (Winter 2019 Edition). https:plato.stanford.edu/archives/win2019/entries/intentionality

James, W. (1900). *Human immortality: Two supposed objections to the doctrine* (2nd ed.). Houghton, Mifflin. (Original work published 1898)

Janet, P. (1889). *L'Automatisme psychologique: Essai de psychologie expérimental sur les forms inférieure de l'activité humaine* [Psychological automatism: An essay of experimental psychology on the inferior forms of human activity]. Félix Alcan.

Kastrup, B. (2019). *The idea of the world: A multidisciplinary argument for the mental nature of reality.* Iff Books.

Kelly, E. F., Kelly, E. W., Crabtree, A., Gauld, A., Grosso, M., & Greyson, B. (2007). *Irreducible mind: Toward a psychology for the 21st century.* Rowman & Littlefield.

Kelly, E. F., Crabtree, A., & Marshall, P. (Eds.) (2015). *Beyond physicalism: Toward reconciliation of science and spirituality.* Rowman & Littlefield.

Kelly, E. F., & Locke, R. G. (2009). *Altered states of consciousness and psi: An historical survey and research prospectus.* Parapsychology Foundation(Original work published 1981)

Kelly, E. F., & Marshall, P. (Eds.) (2021). *Consciousness unbound: Liberating mind from the tyranny of materialism.* Rowman & Littlefield.

Kelly, E. F., & Presti, D. (2015). A psychobiological perspective on “transmission” models. In E. F. Kelly, A. Crabtree & P. Marshall (Eds.), *Beyond physicalism* (pp. 115-155). Lanham, MD: Rowman & Littlefield.

Koch, C. (2012). *Consciousness: Confessions of a romantic reductionist.* MIT Press. https://doi.org/10.7551/mitpress/9367.001.0001

Koch, C. (2019). *The feeling of life itself: Why consciousness is widespread but can’t be computed.* MIT Press. https://doi.org/10.7551/mitpress/11705.001.0001

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nature Reviews Neuroscience, 17*, 307-321. https://doi.org/10.1038/nrn.2016.22

Kripke, S. A. (1980). *Naming and necessity.* Basil Blackwell.

Marshall, P. (2011). The psychical and the mystical: Boundaries, connections, common origins. *Journal of the Society for Psychical Research, 75*(902), 1-13.

Marshall, P. (2021). Mind beyond brain: Surveying the metaphysical landscape. In E. F. Kelly & P. Marshall (Eds.), *Consciousness unbound* (pp. 407-482). Rowman & Littlefield.

Massimini, M., & Tononi, G. (2018). *Sizing up consciousness: Toward an objective measure of the capacity for experience.* Oxford University Press. Doi: 10.1093/oso/9780198728443.001.0001

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring integrated information: Comparison of candidate measures in theory and simulation. *Entropy, 21*(1), 1-30. https://doi.org/10.3390/e21010017

Merker, B., Williford, K., & Rudrauf, D. (2022). The integrated information theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences, 45*, e41:, 1-63. doi:10.1017/S0140525X21080881

Moreira-Almeida, A., & Santos, F. (2013). *Exploring frontiers of the mind-brain relationship.* Springer-Verlag. https://doi.org/10.1007/978-1-4614-0647-1

Muthukumaraswamy, S. D., Carhart-Harris, R. L., Moran, R. J., Brookes, M. J., Williams, T. M., Eritzoe, D., … Nutt, D. J. (2013). Broadband cortical desynchronization underlies the human psychedelic state. *Journal of Neuroscience, 33*(38), 15171-15183. https://doi.org/10.1523/jneurosci.2063-13.2013

Myers, F. W. H. (1903). *Human personality and its survival of bodily death* (2 vols.). Longmans, Green. https://doi.org/10.2307/1412223

Nagel, T. (1979). Brain bisection and the unity of consciousness. In *Mortal Questions* (ch. 11). Cambridge University Press.

Nagel, T. (2012). *Mind and cosmos: Why the materialist neo-Darwinian conception of nature is almost certainly false*. Oxford University Press https://doi.org/10.1093/acprof:oso/9780199919758.001.0001

Nahm, M., Roussseau, D., & Greyson, B. (2017). Discrepancy between cerebral structures and cognitive functioning: A review. *Journal of Nervous and Mental Disease, 205*(12), 967-972. https://doi.org/10.1097/NMD.0000000000000752

Nahm, M., & Weibel, A. (2020). The significance of autoscopies as a time marker for the occurrence of near-death experience. *Journal of Near-Death Studies, 38*(1), 26-50. https://doi.org/10.17514/JNDS-2020-38-1-p26-50

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

Pautz, A. (2019). What is the Integrated Information Theory of consciousness? A catalogue of questions. *Journal of Consciousness Studies, 26*(1-2), 188-215.

Palhano-Fontes, F., Andrade, K. C., Tofoli, L. F., Santos, A. C., Crippa, J. A. S., Hallak, J. E. C., Ribeiros, S., & Araujo, D. B. (2015). The psychedelic state induced by Ayahuasca modulates the activity and connectivity of the Default Mode Network. *PLOS ONE, 10*(2):e0118143. https://doi.org/10.1371/journal.pone.0118143

Pinto, Y., Neville, D. A., Otten, M., Corballis, P. M., Lamme, V. A. F., de Haan, E. H. F., Foschi, N., & Fabri, M. (2017). Split brain: divided perception but undivided consciousness. *Brain, 140*(5), 1231-1237. https://doi.org/10.1093/brain/aww358

Pollan, M. (2018). *How to change your mind: What the new science of psychedelics teaches us about consciousness, dying, addiction, depression, and transcendence.* Penguin.

Popper, K., & Eccles, J. (1977). *The self and its brain: An argument for interactionism.* Springer . https://doi.org/10.1007/978-3-642-61891-8

Presti, D. (2021). Expanding a science of consciousness. In E. F. Kelly & P. Marshall (Eds.), *Consciousness unbound* (pp. 323-358). Rowman & Littlefield.

Prince, M. (1908). *The dissociation of a personality* (2nd ed.). Longmans, Green. (Original work published 1905)

Prince, W. F. (1964). *The case of Patience Worth.* University Books. (Original work published 1927)

Radin, D. I. (2006). *Entangled minds: Extrasensory experiences in a quantum reality.* Simon & Schuster.

Rivas, T., Dirven, A., & Smit, R. H. (2016). *The self does not die: Verified phenomena from near-death studies.* IANDS Publications.

Schaffer, J. (2010). Monism: The priority of the whole. *Philosophical Reviews, 119*(1), 31-76. https://doi.org/10.1215/00318108-2009-025

Schartner, M. M., Carhart-Harris, R. L., Barrett, A. B., Seth, A. K,. & Muthukumaraswamy, S. D. (2017). Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific Reports, 7*(45421). https://doi.org/10.1038/srep46421

Seager, W. (Ed.) (2020). *The Routledge handbook of panpsychism.* Routledge. https://doi.org/10.4324/9781315717708

Searle, J. R. (2013). Can information theory explain consciousness? *The New York Review of Books, 60*(1). https://www.nybooks.com/articles/2013/01/10/can-information-theory-explain-consciousness/

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

Seth, A. K., Izhikevich, E., Reeke, G. N., & Edelman, G. M. (2006). Theories and measures of consciousness: An extended framework. *Proceedings of the National Academy of Sciences USA, 103*(8), 10799-10804. https://doi.org/10.1073/pnas.0604347103

Seth, A. K., Schartner, M., Tagliazucchi, E., Muthukumaraswamy, S., Carhart-Harris, R., & Barrett, A. (2018, October 26). What psychedelic research can and cannot tell us about consciousness. *Scientific American Blog.* https://blogs.scientificamerican.com/observations/what-psychedelic-research-can-and-cannot-tell-us-about-consciousness

Shannon, C. E., & Weaver, W. (1963). *The mathematical theory of communication.* University of Illinois Press.

Strawson, G. (Ed.) (2006). *Consciousness and its place in nature: Does physicalism entail panpsychism?* Imprint Academic.

Swanson, L. R. (2018). Unifying theories of psychedelic drug effects. *Frontiers in Pharmacology, 9,* 1-23. https://doi.org/10.3389/fphar.2018.00172

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(42), https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *Biological Bulletin, 215*(3), 216-242. https://doi.org/10.2307/25470707

Tononi, G. (2012). *PHI: A voyage from the brain to the soul.* Pantheon Books.

Tononi, G. (2015). Integrated information theory. *Scholarpedia, 10,* 4164. https://doi.org/10.4249/scholarpedia.4164

Tononi, G., & Laureys, S. (2009). The neurology of consciousness: An overview. In S. Laureys & G. Tononi (Eds.), *The neurology of consciousness*: *Cognitive neuroscience and neuropathology* (pp. 375-412). Academic Press. https://doi.org/10.1016/B978-0-12-374168-4.00028-9

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B,* 370: 20140167. https://doi.org/10.1098/rstb.2014.0167

Tononi, G. Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*, 450-461. https://doi.org/10.1038/nrn.2016.44

Van Lommel, P., van Wees, R. Meyers, V., & Elfferich, J. (2001). Near-death experiences in survivors of cardiac arrest: A prospective study in the Netherlands. *Lancet, 358*(9298), 2039- 2045. https://doi.org/10.1016/S0140-6736(01)07100-8

Velmans, M. (2021) Is the universe conscious? Reflexive monism and the ground of being. In E. Kelly and P. Marshall (Eds.), *Consciousness unbound* (pp. 175-228). Rowman and Littlefield.

Wackermann, J., & Allefeld, C. (2009). State space representation and global descriptors of brain electrical activity. In C. M. Michel, T. Koenig, D. Brandeis, L. R. R. Gianotti & J. Wackermann (Eds.), *Electrical neuroimaging* (pp. 191-214). Cambridge University Press. https://doi.org/10.1017/CBO9780511596889.010

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*, 440-442. https://doi.org/10.1038/30918

Einige konzeptionelle und empirische Mängel der IIT

Edward F. Kelly

Zusammenfassung: Die Integrierte Informationstheorie des Bewusstseins (IIT) hat innerhalb und außerhalb der wissenschaftlichen Gemeinschaft viel Aufsehen erregt und scheint für viele als der führende Anwärter zu gelten, was eine zufriedenstellende Theorie auf der Grundlage der Systemneurowissenschaften betrifft. Es handelt sich um eine kühne Theorie, die plausible Erklärungen für verschiedene anerkannte neurowissenschaftliche Tatsachen liefert, überraschende Vorhersagen macht, die über die derzeitige wissenschaftliche Orthodoxie hinausgehen, aber potenziell überprüfbar sind, und zur Entwicklung einer scheinbar wirksamen Technik beigetragen hat, die geeignet scheint, das Vorhandensein von Bewusstsein in Organismen nachzuweisen, die nicht in der Lage sind, sich sprachlich auszudrücken wie etwa nichtmenschliche Tiere, Neugeborene und schwer hirngeschädigte Erwachsene. Trotz dieser Vorzüge scheint die IIT grundsätzliche Fehler aufzuweisen: In diesem Beitrag werden zunächst einige wichtige konzeptionelle und technische Fragen aufgegriffen, die bereits früher aufgeworfen wurden, aber noch ungelöst sind - insbesondere Fragen bezüglich des IIT-Konzepts von "Information" und seiner Behandlung des "harten Problems" -, und dann rücken mehrere empirische Phänomene ins Zentrum, die die IIT offenbar nicht zufriedenstellend behandeln kann. Dazu gehören: 1. Fälle von multipler Persönlichkeit oder dissoziativer Identitätsstörung, bei denen komplexe und sich überschneidende Bewusstseinszentren in einzelnen menschlichen Organismen gleichzeitig auftreten; 2. das Scheitern beim Versuch, anzugeben, wie sich die intensive Phänomenologie psychedelischer Zustände direkt in der begleitenden neuroelektrischen Aktivität widerspiegelt; und, was am kritischsten ist, 3. die tiefgreifenden Wirkungen und die die Persönlichkeit verändernden Nahtoderfahrungen (NTEs) unter extremen physiologischen Bedingungen wie beispielsweise Herzstillstand, bei denen der IIT zufolge keinerlei bewusste Erfahrung mehr möglich sein sollte. Diese empirischen Argumente zeigen, dass die IIT selbst unhaltbar ist, und sie gelten auch für ihre physikalistischen Konkurrenten. Es gibt jedoch Alternativen, die wissenschaftlich und philosophisch respektabel sind.

Eberhard Bauer

Algumas insuficiências conceituais e empíricas da TII

Edward F. Kelly

Resumo: A Teoria da Informação Integrada sobre a consciência (TII) tem gerado muita empolgação dentro e fora da comunidade científica, e parece a muitos a principal candidata a uma teoria satisfatória fundamentada em neurociência de sistemas. É uma teoria ousada, que fornece explicações plausíveis para vários fatos neurocientíficos reconhecidos, faz previsões surpreendentes que vão além da ortodoxia científica atual, mas que são potencialmente testáveis, e inspirou o desenvolvimento do que parece ser uma técnica eficaz para detectar a presença da consciência em organismos incapazes de relato verbal, tais como animais não-humanos, recém-nascidos e adultos com o cérebro gravemente prejudicado. Apesar destas virtudes, a TII parece fundamentalmente imperfeita: Este artigo revisita primeiro algumas questões conceituais e técnicas chave que foram levantadas anteriormente, mas que permanecem sem solução - em particular, questões relativas ao conceito de "informação" da TII e sua abordagem do "problema difícil" - e depois se concentra em vários fenômenos empíricos que a TII parece incapaz de lidar de maneira satisfatória. Estes incluem: 1. casos de personalidade múltipla ou transtorno dissociativo de identidade em que centros complexos e sobrepostos de consciência coocorram em organismos humanos únicos; 2. o fracasso da intensa fenomenologia dos estados psicodélicos em se refletir diretamente em atividade neuroelétrica associada; e, de forma mais crítica; 3. a ocorrência de profundas e pessoalmente transformadoras experiências de quase-morte (EQM) sob condições fisiológicas extremas, tais como parada cardíaca, nas quais a TII prevê que nenhuma experiência consciente deveria ser possível. Estes argumentos empíricos mostram que a TII é, em si mesma, insustentável, e se aplicam também a seus concorrentes fisicalistas. Estão disponíveis, no entanto, alternativas cientificas e filosóficas respeitáveis.

Antônio Lima

Algunas deficiencias conceptuales y empíricas de la IIT

Edward F. Kelly

Resumen: La Teoría de Información Integrada de consciencia (IIT) ha generado mucho entusiasmo dentro y fuera de la comunidad científica; a muchos les parece la principal aspirante a una teoría satisfactoria basada en la neurociencia de sistemas. Se trata de una teoría audaz, que proporciona explicaciones plausibles para varios hechos neurocientíficos reconocidos, hace predicciones sorprendentes que van más allá de la ortodoxia científica actual pero que son potencialmente comprobables, y ha inspirado el desarrollo de lo que parece ser una técnica eficaz para detectar la presencia de la consciencia en organismos incapaces de informar verbalmente, incluyendo a animales no humanos, neonatos y adultos con daños cerebrales graves. A pesar de estas virtudes, la IIT parece tener defectos fundamentales: Este artículo discute en primer lugar algunas cuestiones conceptuales y técnicas clave que se han planteado anteriormente pero que siguen sin resolverse

-en particular, cuestiones relativas al concepto de "información" de la IIT y su enfoque sobre el "problema difícil." A continuación, se centra en varios fenómenos empíricos que la IIT parece incapaz de resolver satisfactoriamente incluyendo: 1. casos de personalidad múltiple o de trastorno de identidad disociativo, en los que coexisten centros de consciencia complejos y superpuestos en un solo organismo humano; 2. el hecho de que la intensa fenomenología de los estados psicodélicos no se refleja directamente en la actividad neuroeléctrica que los acompaña; y, lo que es más importante; 3. la ocurrencia de experiencias cercanas a la muerte (NDE), profundas y personalmente transformadoras, en condiciones fisiológicas extremas tales como el paro cardíaco, en las que la IIT predice que no debería ser posible ninguna experiencia consciente. Estos argumentos empíricos muestran que la propia IIT es insostenible y se aplican también a sus teorías competidores fisicalistas. Sin embargo, existen alternativas científica y filosóficamente respetables.

Etzel Cardeña