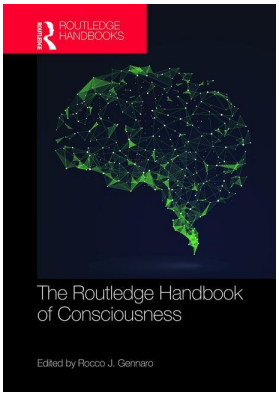

The Routledge Handbook Of Consciousness

Rocco J. Gennaro

Integrated Information Theory

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315676982-11
Francis Fallon
Published online on: 26 Mar 2018

10
INTEGRATED INFORMATION THEORY

Francis Fallon

Integrated Information Theory (IIT) combines Cartesian commitments with claims about engineering that it interprets, in part by citing corroborative neuroscientific evidence, as identifying the nature of consciousness. This borrows from recognizable traditions in the field of consciousness studies, but the structure of the argument is novel.

IIT takes certain features of consciousness to be unavoidably true. Rather than beginning with the neural correlates of consciousness (NCC) and attempting to explain what about these sustains consciousness, IIT begins with its characterization of experience itself, determines the physical properties necessary for realizing these characteristics, and only then puts forward a theoretical explanation of consciousness, as identical to a special case of information instantiated by those physical properties. "The theory provides a principled account of both the quantity and quality of an individual experience… and a calculus to evaluate whether a physical system is conscious" (Tononi and Koch 2015).

1 The Central Claims[1]

IIT takes Descartes very seriously. Descartes located the bedrock of epistemology in the knowledge of our own existence given to us by our thought. "I think, therefore I am" reflects an unavoidable certainty: one cannot deny one's own existence as a thinker (even if one's particular thoughts are in error). For IIT, the relevance of this insight lies in its application to consciousness. Whatever else one might claim about consciousness, one cannot deny its existence.

IIT takes consciousness as primary. What does consciousness refer to here? Before speculating on the origins or the necessary and sufficient conditions for consciousness, IIT gives a characterization of what consciousness means. The theory advances five axioms intended to capture just this. Each axiom articulates a dimension of experience that IIT regards as self-evident. They are as follows:

First, following from the fundamental Cartesian insight, is the axiom of *existence*. Consciousness is real and undeniable; moreover, a subject's consciousness has this reality intrinsically; i.e. it exists from its own perspective.

Second, consciousness has *composition*. In other words, each experience has structure. Color and shape, for example, structure visual experience. Such structure allows for various distinctions.

Third, the axiom of *information*: the way an experience is distinguishes it from other possible experiences. An experience specifies; it is specific to certain things, distinct from others.

Fourth, consciousness has the characteristic of *integration*. The elements of an experience are interdependent. For example, the particular colors and shapes that structure a visual conscious state are experienced together. As we read these words, we experience the font-shape and letter-color inseparably. We do not have isolated experiences of each and then add them together. This integration means that consciousness is irreducible to separate elements. Consciousness is unified.

Fifth, consciousness has the property of *exclusion*. Every experience has borders. Precisely because consciousness specifies certain things, it excludes others. Consciousness also flows at a particular speed.

In isolation, these axioms may seem trivial or overlapping. IIT labels them axioms precisely because it takes them to be obviously true. IIT does not present them in isolation. Rather, they motivate postulates.[2] Each axiom leads to a corresponding postulate identifying a physical property. Any conscious system must possess these properties. The postulates include:

First, the existence of consciousness implies a system of mechanisms with a particular cause-effect power. IIT regards existence as inextricable from causality: for something to exist, it must (be able to) make a difference to other things, and vice versa. (What would it even mean for a thing to exist in the absence of any causal power whatsoever?) Because consciousness exists from its own perspective, the implied system of mechanisms must do more than simply have causal power; it must have *cause-effect power upon itself*.

Second, the compositional nature of consciousness implies that its system's mechanistic elements must have the capacity to *combine*, and that those combinations have cause-effect power.

Third, because consciousness is informative, it must *specify*, i.e. distinguish one experience from others. IIT calls the cause-effect powers of any given mechanism within a system, its cause-effect repertoire. The cause-effect repertoires of all the system's mechanistic elements taken together, it calls its cause-effect structure. This structure, at any given point, is in a particular state. In complex structures, the number of possible states is very high. For a structure to instantiate a particular state is for it to specify that state. The specified state is the particular way that the system is making a difference to itself.

Fourth, consciousness's integration into a unified whole implies that the system must be *irreducible*. In other words, its parts must be interdependent. This in turn implies that every mechanistic element must have the capacity to act as a cause for the rest of the system and to be affected by the rest of the system. If a system can be divided into two parts without affecting its cause-effect structure, it fails to satisfy the requirement of this postulate.

Fifth, the exclusivity of the borders of consciousness implies that the state of a conscious system must be *definite*. In physical terms, the various simultaneous subgroupings of mechanisms in a system have varying cause-effect structures. Of these, only one will have a maximally irreducible cause-effect structure (called the *maximally irreducible conceptual structure*, or MICS). Others will have smaller cause-effect structures, at least when reduced to non-redundant elements. Precisely this – the MICS – is the conscious state.

IIT accepts the Cartesian conviction that consciousness has immediate, self-evident properties, and outlines the implications of these phenomenological axioms for conscious physical systems. This characterization does not exhaustively describe the theoretical ambition of IIT. The ontological postulates concerning physical systems do not merely articulate necessities (or even sufficiencies) for realizing consciousness; the claim is much stronger than this. IIT *identifies* consciousness with a system's having the physical features that the postulates describe. Each conscious state is a MICS, which just is and can only be a system of irreducibly interdependent physical parts whose causal interaction constitutes the integration of information.

An example may help to clarify the nature of IIT's explanation of consciousness. Our experience of a cue ball integrates its white color and spherical shape, such that these elements

are inseparably fused. The fusion of these elements constitutes the structure of the experience: the experience is composed of them. The nature of the experience informs (about whiteness and spherical shape) in a way that distinguishes it from other possible experiences (such as of a blue cube of chalk). This is just a description of the phenomenology of a simple experience (perhaps necessarily awkward, because it articulates the self-evident). Our brain generates the experience through neurons physically communicating with one another, in systems linked by cause-effect power. IIT interprets this physical communication as the integration of information, according to the various constraints laid out in the postulates. The neurobiology and phenomenology converge.

Indeed, according to IIT, the physical state of any conscious system *must* converge with phenomenology; otherwise the kind of information generated could not realize the axiomatic properties of consciousness. We can understand this by contrasting two kinds of information. First, Shannon information: When a digital camera takes a picture of a cue ball, the photodiodes operate in causal isolation from one another. This process does generate information; specifically, it generates observer-relative information. That is, the camera generates the information of an image of a cue ball for anyone looking at that photograph. The information that is the image of the cue ball is therefore relative to the observer; such information is called Shannon information. Because the elements of the system are causally isolated, the system does not make a difference to itself. Accordingly, although the camera gives information to an observer, it does not generate that information for itself. By contrast, consider what IIT refers to as *intrinsic* information: unlike the digital camera's photodiodes, the brain's neurons do communicate with one another through physical cause and effect; the brain does not simply generate observer-relative information, it integrates *intrinsic* information. This information from its own perspective *just is* the conscious state of the brain. The physical nature of the digital camera does not conform to IIT's postulates and therefore does not have consciousness; the physical nature of the brain, at least in certain states, does conform to IIT's postulates, and therefore does have consciousness.

To identify consciousness with such physical integration of information constitutes a bold and novel ontological claim. Again, the physical postulates do not describe one way, or even the best way, to realize the phenomenology of consciousness; the phenomenology of consciousness is one and the same as a system having the properties described by the postulates. It is even too weak to say that such systems "give rise to" or "generate" consciousness. Consciousness is fundamental to these systems in the same way as mass or charge is basic to certain particles.

IIT's conception of consciousness as mechanisms systematically integrating information through cause and effect lends itself to quantification. The more complex the MICS, the higher the level of consciousness: the corresponding metric is *phi*. IIT points to certain cases as illustrating this relation, thereby providing corroborative evidence of its central claims. For example, deep sleep states are less experientially rich than waking ones. IIT predicts, therefore, that such sleep states will have lower *phi* values than waking states. For this to be true, analysis of the brain during these contrasting states would have to show a disparity in the systematic complexity of non-redundant mechanisms. In IIT, this disparity of MICS complexity directly implies a disparity in the amount of conscious integrated information (because the MICS is identical to the conscious state). The neuroscientific findings bear out this prediction.

IIT cites similar evidence from the study of patients with brain damage. For example, we already know that among vegetative patients, there are some whose brain scans indicate that they can hear and process language: when researchers prompt such patients to think about playing tennis, e.g., the appropriate areas of the brain become activated. Other vegetative patients do not respond this way. Naturally, this suggests that the former have a richer degree of consciousness than the latter. When analyzed according to IIT's theory, the former have a higher *phi* metric

than the latter; once again, IIT has made a prediction that receives empirical confirmation. IIT also claims that findings in the analysis of patients under anaesthesia corroborate its claims.

In all these cases, one of two things happens. First, as consciousness fades, cortical activity may become less global. This reversion to local cortical activity constitutes a loss of integration: the system no longer is communicating across itself in as complex a way as it had. Second, as consciousness fades, cortical activity may remain global, but become stereotypical, consisting in numerous redundant cause-effect mechanisms, such that the informational achievement of the system is reduced: a loss of information. As information either becomes less integrated or becomes reduced, consciousness fades, which IIT takes as empirical support of its theory of consciousness as integrated information.

2 Quantifying Consciousness: Measuring *Phi*

IIT strives, among other things, not just to claim the existence of a scale of complexity of consciousness, but to provide a theoretical approach to the precise quantification of the richness of experience for any conscious system. This requires calculating the maximal amount of integrated information in a system: the system's *phi* value can be expressed numerically (at least in principle). It is important to note that not every system with phi has consciousness. A sub- or super-system of an MICS may have phi, but will not have consciousness. A closer look at the digital photography example affords particularly apt illustrations of some of the basic principles involved in quantifying consciousness.

First, a photodiode exemplifies integrated information in the simplest way possible. A photodiode is a system of two elements, which together render it sensitive to two states only: light and dark. After initial input from the environment, the elements communicate input physically with one another, determining the output. So, the photodiode is a two-element system that integrates information. A photodiode not subsumed in another system of greater phi value is the simplest possible example of consciousness.

This consciousness, of course, is virtually negligible. The photodiode's experience of light and dark is not rich in the way that ours is. The level of information of a state depends upon its specifying that state as distinct from others. The repertoire of the photodiodes allows only for the most limited differentiation ('this' vs. 'that'), whereas the repertoire of a complex system such as the brain allows for an enormous amount of differentiation. Even our most basic experience of darkness distinguishes it not only from light, but from shapes, colors, etc.

Second, as noted in Section 1, a digital camera's photodiodes' causal arrangement neatly exemplifies the distinction between integrated and non-integrated information. Putting to one side that each individual photodiode integrates information (as simply as possible), those photodiodes do not take input or give output *to one another*, so the information does not get integrated across the system. For this reason, the camera's image is informative to us, but not to itself. So, each isolated photodiode has integrated information in the most basic way, and would therefore have the lowest possible positive value of phi. The camera's photodiodes taken as a system do not integrate information and have a phi value of zero.

In order to measure the level of consciousness of a system, IIT must describe the amount of its integrated information. This is done by partitioning the system in various ways.[3] If the digital camera's photodiodes are partitioned (say, by dividing the abstract model of its elements in half) no integrated information is lost, because all the photodiodes are in isolation from each other, and so the division does not break any connections. If no logically possible partition of the system results in a loss of connection, the conclusion is that the system does not make a difference to itself. So, in this case, the system has no phi.

Systems with phi will have connections that will be lost by some partitions and not by others. Some partitions will sever from the system elements that are comparatively low in original degree of connectivity to the system; in other words, elements whose (de)activation has few causal consequences upon the (de)activation of other elements. A system where all or most elements have this property will have low phi.

The lack of strong connectivity may be the result of relative isolation, or locality (an element not linking to many other elements, directly or indirectly) or from stereotypicality (where the element's causal connections overlap in a largely redundant way with the causal connection of other elements). A system whose elements are connected more globally and non-redundantly will have higher phi. These descriptions apply, for example, to the cortical activity of sleep and wake states, respectively (see Section 1 above).

A partition that not only separates all elements that do not make a difference to the rest of the system (for reasons of either isolation or redundancy) from those that do make a difference, but also separates those elements whose lower causal connectivity decreases the overall level of integration of the system from those that do not, will thereby have picked out the MICS, which according to IIT is conscious. The degree of that consciousness, its phi, depends upon its elements' level of causal connectivity. This is determined by how much information integration would be lost by the least costly further partition, or, in other words, how much the cause-effect structure of the system would be reduced by eliminating the least causally effective element within the MICS.

3 What IIT's Central Claims Imply

No controversy attaches to the observation that humans experience varying degrees of consciousness. As noted, consciousness decreases during sleep, for example. IIT implies that brain activity during this time will generate either less information or less integrated information, and interprets experimental results concerning cortical activity as bearing this out. By contrast, the cerebellum, which has many neurons, but neurons that are not complexly interconnected and so do not belong to the MICS, does not generate consciousness.

More controversial is the issue of non-human consciousness. IIT counts among its merits that the principles it uses to characterize human consciousness can apply to non-human cases. On IIT, consciousness happens when a system makes a difference to itself at a physical level: elements causally connected to one another in a re-entrant architecture integrate information, and the subset of these with maximal causal power is conscious. The human brain offers an excellent example of re-entrant architecture integrating information, capable of sustaining highly complex MICSs, but nothing in IIT limits the attribution of consciousness to human brains only.

Mammalian brains share similarities in neural and synaptic structure: the human case is not obviously exceptional. Other, non-mammalian species demonstrate behavior associated in humans with consciousness. These considerations suggest that humans are not the only species capable of consciousness. IIT makes a point of remaining open to the possibility that many other species may possess at least some degree of consciousness. At the same time, further study of non-human neuroanatomy is required to determine whether and how this in facts holds true. As mentioned above, even the human cerebellum does not have the correct architecture to generate consciousness, and it is possible that other species have neural organizations that facilitate complex behavior without generating high phi. The IIT research program offers a way to establish whether these other systems are more like the cerebellum or the cerebral cortex in humans. Of course, consciousness levels will not correspond completely to species alone. Within conscious species, there will be a

range of phi levels, and even within a conscious phenotype, consciousness will not remain constant from infancy to death, wakefulness to sleep, and so forth.

IIT claims that its principles are consistent with the existence of cases of dual consciousness within split-brain patients. In such instances, on IIT, two local maxima of integrated information exist separately from one another, generating separate consciousness. IIT does not hold that a system need have only one local maximum, although this may be true of normal brains; in split-brain patients, the re-entrant architecture has been severed so as to create two. IIT also takes its identification of MICSs (through quantification of phi) as a potential tool for assessing other actual or possible cases of multiple consciousness within one brain.

Such claims also allow IIT to rule out instances of aggregate consciousness. The exclusion principle forbids double-counting of consciousness. A system will have various subsystems with phi value, but only the local maxima of phi within the system can be conscious. A normal waking human brain has only one conscious MICS, and even a split-brain patient's conscious systems do not overlap but rather are separate. One's conscious experience is precisely what it is and nothing else. All this implies that, for example, the USA has no superordinate consciousness in addition to the consciousness of its individuals. The local maxima of integrated information reside within the skulls of those individuals; the phi value of the connections among them is much lower.

Although IIT allows for a potentially very wide range of degrees of consciousness and conscious entities, this has its limits. Some versions of panpsychism attribute mental properties to even the most basic elements of the structure of the world, but the simplest conscious entity admitted on IIT to be conscious would have to be a system of at least two elements that have cause-effect power over one another. Otherwise no integrated information exists. Objects such as rocks and grains of sand have no phi (whether in isolation or heaped into an aggregate), and therefore no consciousness.

IIT's criteria for consciousness are consistent with the existence of artificial consciousness. The photodiode, because it integrates information, has a phi value; if not subsumed into a system of higher phi, this will count as local maximum: the simplest possible MICS or conscious system. Many or most instances of phi and consciousness may be the result of evolution in nature, independent of human technology, but this is a contingent fact.

IIT's basic arguments imply, and the IIT literature often explicitly claims, certain important constraints upon artificial conscious systems. Often technological systems involve feed-forward architecture that lowers or possibly eliminates phi, but if the system is physically re-entrant and satisfies the other criteria laid out by IIT, it may be conscious. In fact, according to IIT, we may build artificial systems with a greater degree of consciousness than humans.

At the level of hardware, computation may process information with either feed-forward or re-entrant architecture. In feed-forward systems, information gets processed in only one direction, taking input and giving output. In re-entrant systems, which consist of feedback loops, signals are not confined to movement in one direction only; output may operate as input also.

IIT interprets the integration axiom (the fourth axiom, which says that each experience's phenomenological elements are interdependent) as entailing the fourth postulate, which claims that each mechanism of a conscious system must have the potential to relate causally to the other mechanisms of that system. By definition, in a feed-forward system, mechanisms cannot act as causes upon those parts of the system from which they take input. A purely feed-forward system would have no phi, because although it would process information, it would not integrate that information at the physical level. One implication for artificial consciousness is immediately clear: Feed-forward architectures will not be conscious. Even a feed-forward system that perfectly replicated the behavior of a conscious system would only *simulate* consciousness. Artificial systems will need to have re-entrant structure to generate consciousness.

Furthermore, re-entrant systems may still generate very low levels of phi. Conventional CPUs have transistors that only communicate with several others. By contrast, each neuron of the conscious network of the brain connects with thousands of others, a far more complex re-entrant structure, making a difference to itself at the physical level in such a way as to generate much higher phi value. For this reason, brains are capable of realizing much richer consciousness than conventional computers. The field of artificial consciousness, therefore, would do well to emulate the neural connectivity of the brain.

Still another constraint applies, this one associated with the exclusion (fifth) postulate. A system may have numerous phi-generating subsystems, but according to IIT, only the network of elements with the greatest cause-effect power to integrate information (the maximally irreducible conceptual structure, or MICS) is conscious. Re-entrant systems may have local maxima of phi, and therefore small pockets of consciousness. Those attempting to engineer high degrees of artificial consciousness need to focus their design on creating a large MICS, not simply small, non-overlapping MICSs. If IIT is correct in placing such constraints upon artificial consciousness, deep convolutional networks such as GoogLeNet and advanced projects like Blue Brain may be unable to realize (high levels of) consciousness.

4 Selected Objections

Space prohibits even a cursory description of alternative interpretations of consciousness, as the variety of chapters in this volume alone evidences. Even an exhaustive account of the various objections that have been levelled explicitly at IIT is not possible (nor necessarily desirable) here. What follows will be partial in this sense and in the sense that it reflects the author's opinion of the more serious challenges to IIT.[4]

First, the objection from functionalism: According to functionalism, mental states, including states of consciousness, find explanation by appeal to function. The nature of a certain function may limit the possibilities for its physical instantiation, but the function, and not the material details, is of primary relevance (Dennett 1991, 2005). IIT differs from functionalism on this basic issue: on IIT, the conscious state is identified with the way in which a system embodies the physical features that IIT's postulates describe.

Their opposing views concerning constraints upon artificial consciousness nicely illustrate the contrast between functionalism and IIT. For the functionalist, any system that functions identically to, for example, a conscious human, will by definition have consciousness. Whether the artificial system uses re-entrant or feed-forward architecture is a pragmatic matter. It may turn out that re-entrant circuitry more efficiently realizes the function, but even if the system incorporates feed-forward engineering, so long as the function is achieved, the system is conscious. IIT, on the other hand, expressly claims that a system that performed in a way completely identical to a conscious human, but that employed feed-forward architecture, would only simulate, but not realize consciousness. Put simply, such a system would operate as if it were integrating information, but because its networks would not take output as input, would not actually integrate information at the physical level. The difference would not be visible to an observer, but the artificial system would have no conscious experience.

Those who find functionalism unsatisfactory often take it as an inadequate account of phenomenology: no amount of description of functional dynamics seems to capture, for example, our experience of the whiteness of a cue ball. Indeed, IIT entertains even broader suspicions. Beginning with descriptions of physical systems may never lead to explanations of consciousness. Rather, IIT's approach begins with what it takes to be the fundamental features of consciousness. These self-evident, Cartesian descriptors of phenomenology then lead

to postulates concerning their physical realization; only then does IIT connect experience to the physical.

This methodological respect for Cartesian intuitions has a clear appeal, and the IIT literature largely takes this move for granted, rather than offering outright justification for it. In previous work with Edelman, Tononi discusses machine-state functionalism, an early form of functionalism that identified a mental state entirely with its internal, 'machine' state, describable in functional terms (Edelman and Tononi 2000). Noting that Putnam, machine-state functionalism's first advocate, came to abandon the theory (because meanings are not sufficiently fixed by internal states alone), Tononi rejects functionalism generally. More recently, Koch (2012: 92) describes much work in consciousness as "models that describe the mind as a number of functional boxes," where one box is "magically endowed with phenomenal awareness." (Koch confesses to being guilty of this in some of his earlier work.) He then points to IIT as an exception.

Functionalism is not receiving a full or fair hearing in these instances. Machine-state functionalism is a 'straw man': contemporary versions of functionalism do not commit to an entirely internal explanation of meaning, and not all functionalist accounts are subject to the charge of arbitrarily attributing consciousness to one part of a system. The success or failure of functionalism turns on its treatment of the Cartesian intuitions we all have that consciousness is immediate, unitary, and so on. Rather than taking these intuitions as evidence of the unavoidable truth of what IIT describes in its axioms, functionalism offers a subtle alternative. Consciousness indeed seems to us direct and immediate, but functionalists argue that this 'seeming' can be adequately accounted for without positing a substantive phenomenality beyond function. Functionalists claim that the seeming immediacy of consciousness receives sufficient explanation as a set of beliefs (and dispositions to believe) that consciousness is immediate. The challenge lies in giving a functionalist account of such beliefs: no mean feat, but not the deep mystery that non-functionalists construe consciousness as posing. If functionalism is correct in this characterization of consciousness, it undercuts the very premises of IIT.

These considerations relate to the debate concerning access and phenomenal consciousness. Function may be understood in terms of access. If a conscious system has cognitive access to an association or belief, then that association or belief is conscious. In humans, access is often taken to be demonstrated by verbal reporting, although other behaviors may indicate cognitive access. Functionalists hold that cognitive access exhaustively describes consciousness (Cohen and Dennett 2012). Others hold that subjects may be phenomenally conscious of stimuli without cognitively accessing them. IIT may be interpreted as belonging to the latter category.

Interpretation of the relevant empirical studies is a matter of controversy. The phenomenon known as 'change blindness' occurs when a subject fails to notice subtle differences between two pictures, even while reporting thoroughly perceiving each. Dennett's version of functionalism, at least, interprets this as the subject not having cognitive access to the details that have changed, and moreover as not being conscious of them. The subject overestimates the richness of his or her conscious perception. Certain non-functionalists claim that the subject does indeed have the reported rich conscious phenomenology, even though cognitive access to that phenomenal experience is incomplete. Block (2011), for instance, holds this interpretation, claiming that "perceptual consciousness overflows cognitive access." On this account, phenomenal consciousness may occur even in the absence of access consciousness.

IIT's treatment of the role of silent neurons aligns with the non-functionalist interpretation. On IIT, a system's consciousness grows in complexity and richness as the number of elements that could potentially relate causally within the MICS grows. Such elements, *even when inactive*, contribute to the specification of the integrated information, and so help to fix the phenomenal

nature of the experience. In biological systems, this means that silent but potentially active neurons matter to consciousness.

Such silent neurons are not accessed by the system. According to IIT, these non-accessed neurons still contribute to consciousness. As in Block's non-functionalism, access is not necessary for consciousness. On IIT, it is crucial that these neurons could potentially be active, so they must be accessible to the system. Block's account is consistent with this in that he claims that the non-accessed phenomenal content need not be inaccessible. Koch, separately from his support of IIT, takes the non-functionalist side of this argument in Koch and Tsuchiya (2007); so do Fahrenfort and Lamme (2012); and for a functionalist response to the latter, see Cohen and Dennett (2011, 2012).

Non-functionalist accounts that argue for phenomenal consciousness without access make sense given a rejection of the functionalist claim that phenomenality may be understood as a set of beliefs and associations, rather than a Cartesian, immediate phenomenology beyond such things. If, on the other hand, access can explain phenomenality, then the appeal to silent neurons as – despite their inactivity – having a causal bearing on consciousness, becomes as unmotivated as it is mysterious.

Another important distinction between functionalism and IIT lies in their contrasting ontologies. Functionalist explanations of consciousness do not augment the naturalistic ontology in the way that IIT does. Any account of consciousness that maintains that phenomenal experience is immediately first-personal stands in tension with naturalistic ontology, which holds that even experience in principle will receive explanation without appeal to anything beyond objective, or third-personal, physical features. As noted (see Section 3), among theories of consciousness, those versions of panpsychism that attribute mental properties to basic structural elements depart perhaps most obviously from the standard scientific position. Because IIT limits its attribution of consciousness to particular physical systems, rather than to, for example, particles, it constitutes a somewhat more conservative position than panpsychism. Nevertheless, IIT's claims amount to a radical reconception of the ontology of the physical world.

IIT's allegiance to a Cartesian interpretation of experience from the outset lends itself to a non-naturalistic interpretation, although not every step in IIT's argumentation implies a break from standard scientific ontology. IIT counts among its innovations the elucidation of integrated information, achieved when a system's parts make a difference intrinsically, to the system itself. This differs from observer-relative, or Shannon, information, but by itself stays within the confines of naturalism: for example, IIT could have argued that integrated information constitutes an efficient functional route to realizing states of awareness.

Instead, IIT makes the much bolder claim that such integrated information (provided it is locally maximal) is *identical* to consciousness. The IIT literature is quite explicit on this point, routinely offering analogies to other fundamental physical properties. Consciousness is fundamental to integrated information, in the same way as it is fundamental to mass that space-time bends around it. The degree and nature of any given phenomenal feeling follow basically from the particular conceptual structure that is the integrated information of the system. Consciousness is not a brute property of physical structure per se, as it is in some versions of panpsychism, but it is inextricable from physical systems with certain properties, just as mass or charge is inextricable from (some) particles. So, IIT is proposing a striking addition to what science admits into its ontology.

The extraordinary nature of the claim does not necessarily undermine it, but it may be cause for reservation. One line of objection to IIT might claim that this augmentation of naturalistic ontology is non-explanatory, or even ad hoc. We might accept that biological conscious systems possess neurology that physically integrates information in a way that converges with

phenomenology (as outlined in the relation of the postulates to the axioms), without taking this as sufficient evidence for an identity relation between integrated information and consciousness. In response, IIT advocates might claim that the theory's postulates give better ontological ground than functionalism for picking out systems in the first place.

A second major objection to IIT comes in the form of a reductio ad absurdum argument. The computer scientist Scott Aaronson (2014a) has compelled IIT to admit a counterintuitive implication. Certain systems, which are computationally simple and seem implausible candidates for consciousness, may have values of phi higher even than those of human brains, and would count as conscious on IIT. The IIT response has been to accept the conclusion of the reductio, but to deny the charge of absurdity. Aaronson's basic claim involves applying phi calculation. Advocates of IIT have not questioned Aaronson's mathematics, so the philosophical relevance lies in the aftermath.

IIT refers to richly complex systems such as human brains, or hypothetical artificial systems, in order to illustrate high phi value. Aaronson points out that systems that strike us as much simpler and less interesting will sometimes yield a high phi value. The physical realization of an expander graph (his example) could have a higher phi value than a human brain. A graph has points that connect to one another, making the points vertices and the connections edges. This may be thought of as modelling communication between points. Expander graphs are 'sparse' – having not very many points – but those points are highly connected, and this connectivity means that the points have strong communication with one another. In short, such graphs have the right properties for generating high phi values. Because it is absurd to accept that a physical model of an expander graph could have a higher degree of consciousness than a human being, the theory that leads to this conclusion, IIT, must be false.

Tononi (2014) responds directly to this argument, conceding that Aaronson has drawn out the implications of IIT and phi fairly, even ceding further ground: a two-dimensional grid of logic gates (even simpler than an expander graph) would have a high phi value and would, according to IIT, have a high degree of consciousness. Tononi has already argued that a photodiode has minimal consciousness; to him, accepting where Aaronson's reasoning leads is just another case of the theory producing surprising results. After all, science must be open to theoretical innovation.

Aaronson's rejoinder (2014b) challenges IIT by arguing that it implicitly holds inconsistent views on the role of intuition. In his response to Aaronson's original claims, Tononi disparages intuitions regarding when a system is conscious: Aaronson should not be as confident as he is that expander graphs are not conscious. Indeed, the open-mindedness here suggested seems in line with the proper scientific attitude. Aaronson employs a thought-experiment to draw out what he takes to be the problem. Imagine that a scientist announces that he has discovered a superior definition of temperature and has constructed a new thermometer that reflects this advance. It so happens that the new thermometer reads ice as being warmer than boiling water. According to Aaronson, even if there is merit to the underlying scientific work, it is a mistake for the scientist to use the terms 'temperature' or 'heat' in this way, because it violates what we mean by those terms in the first place: 'heat' means, partly, what ice has less of than boiling water. So, while IIT's phi metric may have some merit, it is not in measuring consciousness degree, because 'consciousness' means, partly, what humans have and expander graphs and logic gates do not have.

One might, in defense of IIT, respond by claiming that the cases are not as similar as they seem, that the definition of heat necessitates that ice has less of it than boiling water and that the definition of consciousness does not compel us to draw conclusions about expander graphs' non-consciousness (strange as that might seem). Aaronson's argument goes further, however,

and it is here that the charge of inconsistency comes into play. Tononi's answer to Aaronson's original reductio argument partly relies upon claiming that facts such as that the cerebellum is not conscious are totally well-established and uncontroversial. (IIT predicts this because the wiring of the cerebellum yields a low phi and is not part of the conscious MICS of the brain.) Here, argues Aaronson, Tononi is depending upon intuition, but it is possible that although the cerebellum might not produce our consciousness, it may have one of its own. Aaronson is not arguing for the consciousness of the cerebellum, but rather pointing out an apparent logical contradiction. Tononi rejects Aaronson's claim that expander graphs are not conscious because it relies on intuition, but here Tononi himself is relying upon intuition. Nor can Tononi here appeal to common sense, because IIT's acceptance of expander graphs and logic gates as conscious flies in the face of common sense.

It is possible that IIT might respond to this serious charge by arguing that almost everyone agrees that the brain is conscious, and that IIT has more success than any other theory in accounting for this, while preserving many of our other intuitions (that animals, infants, certain patients with brain damage, and sleeping adults all have dimmer consciousness than adult waking humans, to give several examples). Because this would accept a certain role for intuitions, it would require 'walking back' the gloss on intuition that Tononi has offered in response to Aaronson's reductio. Moreover, Aaronson's arguments show that such a defense of the overall intuitive plausibility of IIT will face difficult challenges.

5 Conclusion

IIT has a good claim to being the most strikingly original theory of consciousness in recent years. Any attempt to gloss it as a variant of Cartesian dualism, materialism, or panpsychism will obfuscate much more than it illuminates. The efforts of its proponents, especially Tononi and Koch (and their respective research centers) continue to secure its place in the contemporary debate. IIT's novelty notwithstanding, attempts to assess it return us to very familiar ground: its very premises take for granted a highly embattled set of Cartesian principles, and its implications – despite its advocates' protests to the contrary – arguably violate both parsimony and intuition. Its fit with certain empirical evidence suggests that the phi measurement may have scientific utility, but it is far from clear that this implies that IIT has succeeded in identifying the nature of consciousness.

Notes

1 Tononi and Koch (2015) outlines the basics; Oizumi, Albantakis, and Tononi (2014) gives a more technical introduction; see also Tononi (2006, 2008).
2 Tononi (2015) adopts the position that the move from the axioms to the postulates is one of inference to the best explanation, or abduction.
3 This is pragmatically impossible for systems with as many components as the human brain, so an ongoing issue within IIT involves refining approximations of these values.
4 It would be remiss to neglect any mention of Searle's (2013a, 2013b) critique of IIT, but as the response from Koch and Tononi (2013) makes very clear, the objection does not succeed.

References

Aaronson, S. (2014a) "Why I Am Not an Integrated Information Theorist (or, the Unconscious Expander)," [Stable web log post]. May 21. Retrieved from *Shtetl-Optimized*, http://scottaaronson.com/blog. Accessed July 27, 2016.

Aaronson, S. (2014b) "Giulio Tononi and Me: A Phi-nal Exchange," [Stable web log post]. May 30, June 2. Retrieved from *Shtetl-Optimized*, http://scottaaronson.com/blog. Accessed July 27, 2016.
Block, N. (2011) "Perceptual Consciousness Overflows Cognitive Access," *Trends in Cognitive Science* 15: 567–575.
Cohen, M., and Dennett, D. (2011) "Consciousness Cannot be Separated from Function," *Trends in Cognitive Science* 15: 358–364.
Cohen, M., and Dennett, D. (2012) "Response to Fahrenfort and Lamme: Defining Reportability, Accessibility and Sufficiency in Conscious Awareness," *Trends in Cognitive Science* 16: 139–140.
Dennett, D. (1991) *Consciousness Explained*, New York: Little, Brown and Co.
Dennett, D. (2005) *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, London: A Bradford Book, The MIT Press.
Edelman, G., and Tononi, G. (2000) *A Universe of Consciousness: How Matter Becomes Imagination*, New York: Basic Books.
Fahrenfort, J., and Lamme, V. (2012) "A True Science of Consciousness Explains Phenomenology: Comment on Cohen and Dennett," *Trends in Cognitive Science* 16: 138–139.
Koch, C. (2012) *Consciousness: Confessions of a Romantic Reductionist*, Cambridge, MA: The MIT Press.
Koch, C., and Tsuchiya, N. (2007) "Phenomenology without Conscious Access is a Form of Consciousness without Top-Down Attention," *Behavioral and Brain Sciences* 30: 509–510.
Koch, C., and Tononi, G. (2013) "Can a Photodiode be Conscious?" *New York Review of Books* (3/7/13).
Oizumi, M., Albantakis, L. and Tononi, G. (2014) "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," *PLOS Computational Biology* 5: 1–25.
Searle, J. (2013a) "Can Information Theory Explain Consciousness?" *New York Review of Books* (1/10/2013).
Searle, J. (2013b) "Reply to Koch and Tononi," *New York Review of Books* (3/7/13).
Tononi, G. (2008) "Consciousness as Integrated Information: A Provisional Manifesto," *Biology Bulletin* 215: 216–242.
Tononi, G. (2014) "Why Scott Should Stare at a Blank Wall and Reconsider (or, the Conscious Grid)," [Stable web log post]. May 30. Retrieved from *Shtetl-Optimized*, http://scottaaronson.com/blog. Accessed July 27, 2016.
Tononi, G. (2015) "Integrated Information Theory," *Scholarpedia* 10: 4164. http://www.scholarpedia.org/w/index.php?title=Integrated_information_theory&action=cite&rev=147165. Accessed June 27, 2016.
Tononi, G., and Koch, C. (2015) "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society, Philosophical Transactions B* 370, DOI:10.1098/rstb.2014.0167.

Related Topics

Materialism
Dualism
Idealism, Panpsychism, and Emergentism
Biological Naturalism and Biological Realism
Robot Consciousness
Animal Consciousness