Integrated Information Theory of Consciousness

Integrated Information Theory (IIT) offers an explanation for the nature and source of consciousness. Initially proposed by Giulio Tononi in 2004, it claims that consciousness is identical to a certain kind of information, the realization of which requires physical, not merely functional, integration, and which can be measured mathematically according to the *phi* metric.

The theory attempts a balance between two different sets of convictions. On the one hand, it strives to preserve the Cartesian intuitions that experience is immediate, direct, and unified. This, according to IIT's proponents and its methodology, rules out accounts of consciousness such as functionalism that explain experience as a system operating in a certain way, as well as ruling out any eliminativist theories that deny the existence of consciousness. On the other hand, IIT takes neuroscientific descriptions of the brain as a starting point for understanding what must be true of a physical system in order for it to be conscious. (Most of IIT's developers and main proponents are neuroscientists.) IIT's methodology involves characterizing the fundamentally subjective nature of consciousness and positing the physical attributes necessary for a system to realize it.

In short, according to IIT, consciousness requires a grouping of elements within a system that have physical cause-effect power upon one another. This in turn implies that only reentrant architecture consisting of feedback loops, whether neural or computational, will realize consciousness. Such groupings make a difference to themselves, not just to outside observers. This constitutes integrated information. Of the various groupings within a system that possess such causal power, one will do so maximally. This local maximum of integrated information is identical to consciousness.

IIT claims that these predictions square with observations of the brain's physical realization of consciousness, and that, where the brain does not instantiate the necessary attributes, it does not generate consciousness. Bolstered by these apparent predictive successes, IIT generalizes its claims beyond human consciousness to animal and artificial consciousness. Because IIT identifies the subjective experience of consciousness with objectively measurable dynamics of a system, the degree of consciousness of a system is measurable in principle; IIT proposes the phi metric to quantify consciousness.

Table of Contents

1. The Main Argument

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).

a. Cartesian Commitments

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.

i. Axioms

IIT takes consciousness as primary. 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.

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; 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 is 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.

ii. Postulates

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. Sometimes the IIT literature refers to phenomenological axioms and ontological postulates. Each axiom leads to a corresponding postulate identifying a physical property. Any conscious system must possess these properties.

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, or 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 on 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. This is 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 is the conscious state.

b. The Identity of Consciousness

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 maximally irreducible conceptual structure, 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 us 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.

Theories of consciousness need to account for what is sometimes termed the "binding problem." This concerns the unity of conscious experience. Even a simple experience like viewing a cue ball unites different elements such as color, shape, and size. Any theory of consciousness will need to make sense of how this happens. IIT's account of the integration of information may be understood as a response to this problem.

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, there is 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 an ontological claim. 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.

i. Some Predictions

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. Sometimes the IIT literature uses the term "prediction" to refer to implications of the theory whose falsifiability is a matter of controversy. This section will focus on more straightforward cases of prediction, where the evidence is consistent with IIT's claims. These cases provide corroborative evidence that enhance the plausibility of IIT.

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. On 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, 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 anesthesia 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.

c. Characterizing the Argument

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. While IIT's proponents strive for clarity in the exposition of their work by breaking it down into the simpler elements of axioms, propositions, and identity claims, the nature of the relations between these parts remains largely implicit in the IIT literature. To evaluate the explanatory success or failure of IIT, it should prove helpful to attempt an explication of these logical relations. This requires characterizing the relationship of the axioms with the postulates, and of the identity claims with the axioms, postulates, and supporting neuroscientific evidence.

The axioms, of course, count as premises. These premises seem to lead to the postulates: each postulate flows from a corresponding axiom. At the same time, IIT describes these postulates as unproven assumptions, which seems at odds with their being conclusions drawn from the axioms. Consider the first axiom and its postulate, concerning existence. The axiom states that consciousness exists, and more specifically, exists intrinsically; the postulate holds that this requires a

conscious system to have cause-effect power, and more specifically, to have this power over itself. The link involves, in part, the claim that existence implies cause-effect power. This claim that for a thing to exist, it must be able to make a difference, is plausible, but not self-evident. Nor does the axiomatic premise alone deductively imply this postulate. Epiphenomenalists, for example, claim that conscious mental states, although existent and caused, do not cause further events; they do not make a difference. Epiphenomenalists certainly do not go on to identify consciousness with physically causal systems, as IIT does.

Tononi (2015) adopts the position that the move from the axioms to the postulates is one of inference to the best explanation, or abduction. On this line, while the axioms do not deductively imply the postulates, the postulates have more than mere statistical inductive support. For example, consider the observation that human brains, which on IIT are conscious systems, have cause-effect power over themselves. Minimally, this offers a piece of inductive support for describing conscious systems in general as having such a power. Tononi takes a stronger line, claiming that a system's property of having cause-effect power over itself most satisfyingly explains its intrinsic existence. So, what makes the brain a system at all, capable of having its own consciousness, is its ability to make a difference to itself. This illustrates the relation of postulates such as the first, concerning cause-effect power, to axioms such as the first axiom, concerning intrinsic existence, by appeal to something like explanatory fit, or satisfactoriness, which is to characterize that relation abductively.

In any case, IIT moves from the sub-conclusion about the postulates to a further conclusion, the identity claim: consciousness is identical to a system's having the physical properties laid out by the postulates, which realize the phenomenology described by the axioms. Here again, the abductive interpretation remains an option. On this interpretation, the *conjunction* of the physical features of the postulates provides the most satisfactory explanation for the identity of consciousness.

This breakdown of the argument reveals the separability of the two parts. A less ambitious version of IIT might have limited itself to the first part, claiming that the physical features described by the postulates are the actual and/or best ways of realizing consciousness, or more strongly that they are necessary and/or sufficient, without going on to say that consciousness is identical to a system having these properties. The foregoing paragraphs outlined the possible motivation for the identification claim as lying in the abductive interpretation.

The notion of best explanation is notoriously slippery, but also ubiquitous in science. From an intuitive point of view one might regard the content of the conjunction of the postulates as apt for accounting for the phenomenology, but

one might, motivated by theoretical conservatism, stop short of describing this as an identity relation. One clue as to why IIT does not take this tack may lie in IIT's methodological goal of parsimony, something the literature mentions with some regularity. Perhaps the simplicity of identifying consciousness with a system's having certain intuitively apt physical properties outweighs the non-conservatism of the claim that consciousness is fundamental to such systems the way mass is fundamental to a particle.

2. The *Phi* Metric

a. The Main Idea

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. IIT refers to this as the system's phi value, which can be expressed numerically, at least in principle.

Digital photography 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, and so forth.

Second, 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.

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.

IIT attributes consciousness to certain systems, or networks. These can be understood abstractly as models. A computer's hardware may be modelled by logic circuits, which represent its elements and their connections as interconnected logic gates. The way a particular connection within this network mediates input and output determines what kind of logic gate it is. For example, consider a connection that takes two inputs, either which can be True or False, and then gives one output, True or False. The AND logic gate would give an output of True if both inputs were True or if both inputs were False. In other words, if both the one AND the other have the same value, the AND gate gives a True output. Such modelling captures the dynamics of binary systems, with "True" corresponding to 1 and "False" to 0. The arrangement of a network's various logic gates (which include not only AND, but also OR, NOT, XOR, among others) determines how any particular input to the system at time-step 1 will result in output at time-step 2, and so on for that system. The brain can be modelled this way too. Input can come from a prior brain state, or from other parts of the nervous system, through the senses. The input causes a change in brain state, depending on the organization of the particular brain, which can be articulated in abstract logic.

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. 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 of interest to IIT 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 it could be 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. 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 maximally irreducible conceptual structure (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.

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.

If we were to take a non-MICS subsystem of a network, which in isolation still has causal power over itself, articulable as a logic circuit, then that would have phi. Were it indeed in isolation, it would have its own MICS, and its phi would correspond to that system's degree of consciousness. It is, however, not in isolation, but rather part of a larger system.

IIT interprets the exclusion axiom—that any conscious system is in one conscious state only, excluding all others—as implying a postulate that holds that, at the level of the physical system, there be no "double counting" of consciousness. So, although a system may have multiple subsystems with phi, only the MICS is conscious, and only the phi value of the MICS (sometimes called phi max) measures conscious degree. The other phi values measure degrees of non-conscious integrated information. So, for example, each of a person's visual cortices does not enjoy its own consciousness, but parts of each belong to a single MICS, which is the person's one unitary consciousness.

If we were to take a supersystem of an MICS, one that includes the MICS and also other associated elements with lower connectivity, we could again assign it a phi value, but this would not measure the local maximum of integrated information. The supersystem integrates information, but not maximally, and its phi is therefore not a measure of consciousness. This is probably best understood by example: a group of people in a discussion integrate information, but the connective degree among them is lower than the degree of connectivity within each individual. The group as such has no consciousness, but each individual person—or, more properly, the MICS of each—does. The individuals' MICSs are local maxima of integrated information and therefore conscious.

b. Some Issues of Application

The number of possible partitions of a system, called Bell's number, grows immensely as the number of elements increases. For example, the tiny nematode, a simple species of worm, has 302 neurons, "and the number of ways that this network can be cut into parts is the hyperastronomical 10 followed by 467 zeros" (Koch, 2012). Calculating phi precisely for much more complex systems such as brains eludes computation pragmatically, although not in principle. In the absence of precise phi computation, IIT employs mathematical "heuristics, shortcuts, and approximations" (ibid.). The IIT literature includes several different mathematical interpretations of phi calculation, each intended to replace the last; it is not yet clear that IIT has a settled account of it. Proponents of IIT hold that the mathematical details will enable the application, but not bear on the merits, of the deeper theoretical claims. At least one serious objection to IIT, however, attempts a *reductio ad absurdum* of those deeper claims via analysis of the mathematical implications.

It is clear that, whatever the mathematical details, the basic principles of phi imply that biological nervous systems such as the brain will be capable of having very high phi, because neurons often have thousands of connections to one another. On the other hand, a typical circuit in a standard CPU only makes a few connections to other circuits, limiting the potential phi value considerably. It is also clear that even simple systems will have at least some phi value, and that, provided they constitute local maxima, will have a corresponding measure of consciousness.

IIT does not intend phi as a measure of the quality of consciousness, only of its quantity. Two systems may have the same phi value but different MICS organizations. In this case, each would be conscious to the same degree, but the nature of the conscious experience would differ. The phi metric captures one dimension, the amount of integrated information, of a system. IIT does address the quality of consciousness abstractly, although not with phi. A system's model includes its elements and their connections, whose logic can be graphed as a constellation of (de)activated points with lines between them representing (de)activated connections. This is, precisely, its conceptual structure. Recall that the maximally irreducible conceptual structure, or MICS, is, on IIT, conscious. A graph of the MICS, according to IIT, captures its unique shape in quality space, the shape of that particular conscious state. In other words, this is the abstract form of the quality of the experience, or the experience's form "seen from the outside." The perspective "from the inside" is available only to the system itself, whose making differences to itself intrinsically, integrating information in one of many possible forms, just precisely is its experience. The phenomenological nature of the experience, its "qualia," are evident only from the perspective of the conscious

system, but the logical graph of its structure is a complete representation of its qualitative properties.

3. Situating the Theory

a. Some Prehistory

IIT made its explicit debut in the literature in 2004, but has roots in earlier work. Giulio Tononi, the theory's founder and major proponent, worked for many years with Gerald Edelman. Their work rejected the notion that mental events such as consciousness will ever find full explanation by reference to the functioning of a system. Such functionalism, to them, ignores the crucial issue of the physical substrate itself. They especially emphasized the importance of re-entry. To them, only a system composed of feedback loops, where input may also serve as output, can integrate information. Feed-forward systems, then, do not integrate information. Even before the introduction of IIT, Tononi was claiming that integrated information was essential to the creation of a scene in primary consciousness.

Christof Koch, now a major proponent of IIT, collaborated for a long time with Francis Crick. Much of their earlier work focused on identifying the neural correlates of consciousness (NCC), especially in the visual system. While such research advances particular knowledge about the mechanisms of one set of conscious states, Crick and Koch came to see this work as failing to address the deeper problems of explaining consciousness generally. Koch also rejects the idea that identifying the functional dynamics of a system aptly treats what makes that system conscious. He came to regard information theory as the correct approach for explaining consciousness.

So, the two thinkers who became IIT's chief advocates arrived at that position after close neuroscientific research with Nobel Laureates who eschewed functional approaches to consciousness, favoring investigation of the relation of physical substrate to information generation.

b. IIT's Additional Support

As of 2016, Tononi, IIT's creator, runs the Center for Sleep and Consciousness at the University of Madison-Wisconsin. The Center has more than forty researchers, many of whom work in its IIT Theory Group. Koch, a major supporter of IIT, heads the prestigious Allen Institute for Brain Science. The Institute has links with the White House Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative, as well as the European Human Brain Project (HBP). IIT's body of literature continues to grow, often in the form of publications associated with the Center and the Institute. The public reputation of these organizations, as well as of Tononi and Koch's earlier work, lends a certain authority or celebrity to IIT. The

theory has enjoyed ample attention in mainstream media. Nevertheless, IIT remains a minority position among neuroscientists and philosophers.

c. IIT as *Sui Generis*

IIT does not fit neatly into any other school of thought about consciousness; there are points of connection to and departures from many categories of consciousness theory.

IIT clearly endorses a Cartesian interpretation of consciousness as immediate; such an association is unusual for a self-described naturalistic or scientific theory of consciousness. Cartesian convictions do inform IIT's axioms and so motivate its overall methodological approach. To label IIT as a Cartesian theory generally, however, would be misleading. For one thing, like most modern theories of consciousness, it dissociates itself from the idea of a Cartesian theatre, or single point in the brain that is the seat of consciousness. Moreover, it is by no means clear how IIT stands in relation to dualism. Certainly, IIT does not advertise itself as positing a mental substance that is separate from the physical. At the same time, it draws an analogy between its identification of consciousness as a property of certain integrated information systems and physics' identification of mass or charge as a property of particles. One might interpret such introduction of immediate experience into the naturalistic ontology as having parallels with positing a new kind of mental substance. The literature will occasionally describe IIT as a form of materialism. It is true that IIT theorists do focus on the material substrate of informational systems, but again, one might challenge whether a theory that asserts direct experience as fundamental to substrates with particular architectural features is indeed limiting itself to reference to material in its explanation.

In describing the features of conscious systems, IIT will make reference to function, but IIT rejects functionalism outright. To IIT theorists, articulating the functional dynamics of a system alone will never do justice to the immediate nature of experience.

d. Relation to Panpsychism

The IIT literature, not only from Tononi but also from Koch and others, refers with some regularity to panpsychism—broadly put, any metaphysical system that attributes mental properties to basic elements of the world—as sharing important ground with IIT. Panpsychism comes in different forms, and the precise relationship between it and IIT has yet to be established. Both IIT and panpsychism strongly endorse Cartesian commitments concerning the immediate nature of experience. IIT, however, only attributes mental properties to re-entrant architectures, because it claims that only these will integrate information; this is

inconsistent with any version of panpsychism that insists upon attributing mental properties to even more basic elements of the structure of existence.

i. Relation to David Chalmers

Of the various contemporary philosophical accounts of consciousness, IIT intersects perhaps most frequently with the work of David Chalmers. This makes sense, not only given Chalmers's panpsychist leanings, but also given the express commitment of both to a Cartesian acceptance of the immediacy of experience, and a corresponding rejection of functionalist attempts to explain consciousness.

Moreover, Chalmers's discussion of the relation of information to consciousness strongly anticipates IIT. Before the introduction of IIT, Chalmers had already endorsed the view of information as involving specification, or reduction of uncertainty. IIT often echoes this, especially in connection with the third axiom and postulate. Chalmers also characterizes information as making a difference, which relates to IIT's first postulate especially. These notions of specification and difference-making are familiar from the standard Shannon account of information, but the point is that both Chalmers and IIT choose to understand consciousness partly by reference to these concepts rather than to information *about* something.

A major theme in Chalmers's work involves addressing the problem of the precise nature of the connection between physical systems and consciousness. Before IIT, Chalmers speculated that information seems to be the connection between the two. If this is the case, then phenomenology is the realization of information. Chalmers suggests that information itself may be primitive in the way that mass or charge is. Now, this does not directly align with IIT's later claims of how consciousness is fundamental to certain informational systems in the same way that mass or charge is fundamental to particles, but the parallels are clear enough. Similarly, Chalmers's description of a "minimally sufficient neural system" as the neural correlate of consciousness resembles IIT's discussion of the MICS. Both also use the term "core" in this context. It comes as no surprise that IIT and Chalmers have intersected when we read Chalmers's earlier claims: "Perhaps, then, the intrinsic nature required to ground the information states is closely related to the intrinsic nature present in phenomenology. Perhaps one is even constitutive of the other" (Chalmers, 1996).

Still, the relationship between Chalmers's work and IIT is not one of simple alliance. Despite the apparent similarity of their positions on what is fundamental, there is an important disagreement. Chalmers takes the physical to derive from the informational, and grounds the realization of phenomenal space—the instantiation of conscious experience—not upon causal "differences that make a difference," but upon the intrinsic qualities of and structural relations among experiences. IIT regards consciousness as being intrinsic to certain causal

structures, which might be read as the reverse of Chalmers's claim. In describing his path to IIT, Koch endorses Chalmers as "a philosophical defender of information theory's potential for understanding consciousness" while faulting Chalmers's work for not addressing the internal organization of conscious systems (Koch, 2012). To Koch, treating the architecture is necessary because consciousness does not alter in simple covariance with change in amounts of bits of information. Because IIT addresses the physical organization, it struck Koch as superior.

There has been some amount of cross-reference between IIT and Chalmers in the literature, although important differences are apparent. For example, Chalmers famously discusses the possibility of a zombie, or operating physical match of a human that does not have experience. On IIT, functional zombies are possible, but not zombies whose nervous connections duplicate our own. In other words, if a machine were built to imitate the behavior of a human perfectly, but whose hardware involved feed-forward circuits, then it would generate possibly no phi, or more likely low, local phi, rather than the high phi of human consciousness. But if we posit that the machine replicates the connections of the human down to the level of the hardware, then it would follow that the system would integrate the same level of phi and would be equally conscious.

Chalmers (2016) writes that IIT "can be construed as a form of emergent panpsychism," which is true in a sense, but requires qualification. By "emergent panpsychism" Chalmers means that IIT posits consciousness as fundamental not to the merest elements of existence but to certain structures that emerge at the level of particles' relations to one another. This is a fair assessment, whether or not IIT's advocates choose to label the theory this way. But in the difference between emergent and non-emergent lies the substance of IIT: what precisely makes one structure "emerge" conscious and another not is what IIT hopes to explain. Non-emergent panpsychism of the sort associated with Chalmers by definition pitches its explanation at a different level. Indeed, it does not necessarily grant the premise that there exist non-conscious elements, let alone structures, in the first place. Despite the similarities between IIT and some of Chalmers's work, the two should not be confused.

4. Implications

a. The Spectrum of Consciousness

It is widely accepted that humans experience varying degrees of consciousness. In sleep, for example, the richness of experience diminishes, and sometimes we do not experience at all. IIT implies that brain activity during this time will generate either less information or less integrated information, and interprets experimental results as bearing this out. On IIT, the complexity of physical connections in the

MICS corresponds to the level of consciousness. 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.

There does not exist a widely accepted position on 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 United States of America 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 so 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. 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.

b. IIT and Physics

IIT has garnered some attention in the physics literature. Even if one accepts the basic principles of IIT, it still remains open to offer a different account of the physical particulars. Tononi and the other proponents of IIT coming from neuroscientific backgrounds tend to offer description at a classical grain. They frame integrated information by reference to neurons and synapses in the case of brains, and to re-entrant hardware architecture in the case of artificial systems. Such descriptions stay within the classical physics paradigm. This does not exhaust the theoretical problem space for characterizing integrated information.

One alternative (Barrett, 2014) proposes that consciousness comes from the integration of information intrinsic to fundamental fields. This account calls for reconceiving the phi metric, which in its 2016 form applies only to discrete systems, and not to electromagnetic fields. Another account (Tegmark, 2015) also

proposes non-classical physical description of conscious, integrated information. This generalizes beyond neural or neural-type systems to quantum systems, suggesting that consciousness is a state of matter, whimsically labelled "perceptronium."

c. Artificial Consciousness

IIT's basic arguments imply, and the IIT literature often explicitly claims, certain important constraints upon artificial conscious systems.

i. Constraints on Structure/Architecture

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 would 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 fifth postulate, the postulate of exclusion. 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.

ii. Relation to "Silent Neurons"

IIT's third postulate has a somewhat counterintuitive implication. The third axiom claims that each conscious experience is precisely what it is; that is, it is distinct from other experiences. The third postulate claims that, in order to realize this feature of consciousness, a system must have a range of possible states, describable by reference to the cause-effect repertoires of its mechanistic elements. The system realizes a specific conscious state by instantiating one of those particular physical arrangements.

An essential component of the phenomenology of a conscious state is the degree of specificity: a photodiode that registers light specifies one of only two possible states. IIT accepts that such a simple mechanism, if not subsumed under a larger MICS, must be conscious, but only to a negligible degree. On the other hand, when a human brain registers light, it distinguishes it from countless other states; not only from dark, but from different shades of color, from sound, and so forth. The brain state is correspondingly more *informative* than the photodiode state.

This means that not only active neuronal firing, but also neuronal silence, determines the nature of a conscious state. Inactive parts of the complex, as well as active ones, contribute to the specification at the physical level, which IIT takes as the realization of the conscious state.

It is important not to conflate silent, or inactive, neurons of this kind with de-activated neurons. Only neurons that genuinely fall within the cause-effect repertoires of the mechanistic elements of the system count as contributing to specification, and this applies to inactive as well as to active ones. If a neuron was incapable all along of having causal power within the MICS, its inactivity plays no role in generating phenomenology. Likewise, IIT predicts that should neurons otherwise belonging to cause-effect repertoires of the system be rendered incapable of such causation (for example, by optogenetics), their inactivity would not contribute to phenomenology.

5. Objections

a. The Functionalist Alternative

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. 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.

i. Rejecting Cartesian Commitments

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 (2000), 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. 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) 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 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.

ii. Case Study: Access *vs.* 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.

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. 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 (Koch and Tsuchiya, 2007); so do Fahrenfort and Lamme (2012); 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.

iii. Challenging IIT's Augmentation of Naturalistic Ontology

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. 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 stronger 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 an 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.

b. Aaronson's *Reductio ad Absurdum*

The computer scientist Scott Aaronson (on his blog *Shtetl-Optimized*; see Horgan (2015) for an overview) 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. Aaronson's argument is intended as a *reductio ad absurdum*; the IIT response has been to accept its conclusion, 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 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.

c. Searle's Objection

In one of very few published discussions of IIT by a philosopher, John Searle (2013a) has come out against it, criticizing its emphasis on information as a departure from the more promising "biological approach." His objections may be divided into two parts; Koch and Tononi (2013) have offered a response.

First, Searle claims that in identifying consciousness with a certain kind of information, it has abandoned causal explanation. Appeal to cause should be the proper route for scientific explanation, and Searle maintains, as he has throughout his career, that solving the mystery of consciousness will depend upon the explication of the causal powers special to the brain that give rise to experience. Information fails aptly to address the problem because it is observer-relative; indeed, it is relative to the conscious observer. A book or computer, to take typical examples of objects associated with information, does not contain information except insofar as it is endowed by the conscious subject. Information is in the eye of the beholder. The notion of information presupposes consciousness, rather than explaining it.

Second, according to Searle, IIT leads to an absurdity, namely panpsychism, which is sufficient reason to reject it. He interprets IIT as imputing consciousness to all systems with causal relations, and so it follows on IIT that consciousness is

"spread over the universe like a thin veneer of jam." A successful theory of consciousness will have to appreciate that consciousness "comes in units" and give a principled account of how and why this is the case.

Koch and Tononi's response (2013) addresses both strands of Searle's argument. First, they agree that Shannonian information is observer-relative, but point out that integrated information is non-Shannonian. IIT defines integrated information as necessarily existing with respect to itself, which they understand in expressly causal terms, as a system whose parts make a difference to that system. Integrated information systems therefore exist intrinsically, rather than relative to observers. Not only does IIT attend to the observer-relativity point, then, but also does so in a way that, contrary to Searle's characterization, crucially incorporates causality.

Second, they deny that IIT implies the kind of panpsychism that Searle rejects as absurd. As they point out, IIT only attributes consciousness to local maxima of integrated information (MICS), and although that implies that some simple systems such as the isolated photodiode have a minimal degree of consciousness, it provides a principle to determine which "units" are conscious, and which are not. As Tononi had already put it, before Searle's charges: "How close is this position to panpsychism, which holds that everything in the universe has some kind of consciousness? Certainly, the IIT implies that many entities, as long as they include some functional mechanisms that can make choices between alternatives, have some degree of consciousness. Unlike traditional panpsychism, however, the IIT does not attribute consciousness indiscriminately to all things. For example, if there are no interactions, there is no consciousness whatsoever. For the IIT, a camera sensor as such is completely unconscious…" (Tononi, 2008).

Although Searle offers a rejoinder (2013b) to Tononi and Koch's response, it largely rehearses the original claims. Regardless of whether IIT is true, Tononi and Koch have given good reason to read it as addressing precisely the concerns that Searle raises. Arguably, then, Searle might have reason to embrace IIT as a theory of consciousness that at least attempts a principled articulation of the special casual powers of the brain, which Searle has regarded for many years as the proper domain for explaining consciousness.

6. References and Further Reading

- Barrett, Adam. "An Integration of Integrated Information Theory with Fundamental Physics." *Frontiers in Psychology*, 5 (63). 2014.
    - Calls for a re-conception of phi with respect to electromagnetic fields.

- Block, Ned. “Perceptual Consciousness Overflows Cognitive Access.” *Trends in Cognitive Science*, 15 (12). 2011.
    - Argues for the distinction between access and phenomenal consciousness.
- Chalmers, David. *The Conscious Mind*. New York: Oxford University Press. 1996.
    - A major work, relevant here for its discussions of information, consciousness, and panpsychism.
- Chalmers, David. “The Combination Problem for Panpsychism.” In L. Jaskolla and G. Bruntup (Ed.s) *Panpsychism*. Oxford University Press. 2016.
    - Updated take on a classical problem; Chalmers makes reference to IIT here.
- Cohen, Michael, and Daniel Dennett. “Consciousness Cannot be Separated from Function.” *Trends in Cognitive Science*, 15 (8). 2011.
    - Argues for understanding phenomenal consciousness as access consciousness.
- Cohen, Michael, and Daniel Dennett. “Response to Fahrenfort and Lamme: Defining Reportability, Accessibility and Sufficiency in Conscious Awareness.” *Trends in Cognitive Science*, 16 (3). 2012.
    - Further defends understanding phenomenal consciousness as access consciousness.
- Dennett, Daniel. *Consciousness Explained*. Little, Brown and Co. 1991.
    - Classic, comparatively accessible teleofunctionalist account of consciousness.
- Dennett, Daniel. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. London, England: The MIT Press. 2005.
    - A concise but wide-ranging, updated defense of functionalist explanation of consciousness.
- Edelman, Gerald. *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books. 1989.
    - Influential upon Tononi’s early thinking.
- Edelman, Gerald, and Giulio Tononi. *A Universe of Consciousness: How Matter Becomes Imagination*. New York: Basic Books. 2000.

    - Puts forward many of the arguments that later constitute IIT.
- Fahrenfort, Johannes, and Victor Lamme. “A True Science of Consciousness Explains Phenomenology: Comment on Cohen and Dennett.” *Trends in Cognitive Science*, 16 (3). 2012.
    - Argues for the access/phenomenal division supported by Block.
- Horgan, John. “Can Integrated Information Theory Explain Consciousness?” *Scientific American*. 1 December 2015. http://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/
    - Gives an overview of an invitation-only workshop on IIT at New York University that featured Tononi, Koch, Aaronson, and Chalmers, among others.
- Koch, Christof. *Consciousness: Confessions of a Romantic Reductionist*. The MIT Press. 2012.
    - Intellectual autobiography, in part detailing the author’s attraction to IIT.
- Koch, Christof, and Naotsugu Tsuchiya. “Phenomenology Without Conscious Access is a Form of Consciousness Without Top-down Attention.” *Behavioral and Brain Sciences,* 30 (5-6) 509-10. 2007.
    - Also argues for the access/phenomenal division supported by Block.
- Koch, Christof, and Giulio Tononi. “Can a Photodiode Be Conscious?” *New York Review of Books,* 7 March 2013.
    - Responds to Searle’s critique.
- Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. “From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0.” *PLOS Computational Biology*, 10 (5). 2014. Doi: 10.1371/journal.pcbi.1003588.
    - Technically-oriented introduction to IIT.
- Searle, John. “Minds, brains and programs.” In Hofstadter, Douglas and Daniel Dennett, (Eds.). *The Mind’s I: Fantasies and Reflections on Self and Soul* (pp. 353-373). New York: Basic Books. 1981.
    - Searle’s classic paper on intentionality and information.
- Searle, John. *The Rediscovery of Mind*. Cambridge, MA: The MIT Press. 1992.

    - Fuller explication of Searle's views on intentionality and information.
- Searle, John. "Can Information Theory Explain Consciousness?" *New York Review of Books* 10 January 2013(a).
    - Objects to IIT.
- Searle, John. "Reply to Koch and Tononi." *New York Review of Books*. 7 March 2013(b).
    - Rejoinder to Koch and Tononi's response to his objections to IIT.
- Tegmark, Max. "Consciousness as a State of Matter." *Chaos, Solitons & Fractals*. 2015.
    - Proposes a re-conception of phi, at the quantum level.
- Tononi, Giulio. "An Information Integration Theory of Consciousness." *BMC Neuroscience,* 5:42. 2004.
    - The earliest explicit introduction to IIT.
- Tononi, Giulio. "Consciousness as Integrated Information: A Provisional Manifesto." *Biology Bulletin,* 215: 216–242. 2008.
    - An early overview of IIT.
- Tononi, Giulio. "Integrated Information Theory." *Scholarpedia*, 10 (1). 2015. http://www.scholarpedia.org/w/index.php?title=Integrated_information_theory&action=cite&rev=147165.
    - A thorough synopsis of IIT.
- Tononi, Giulio, and Gerald Edelman. "Consciousness and Complexity." *Science*, 282 (5395). 1998.
    - Anticipates some of IIT's claims.
- Tononi, Giulio, and Christof Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society, Philosophical Transactions B,* 370 (1668). 2015. Doi: 10.1098/rstb.2014.0167
    - Perhaps the most accessible current introduction to IIT, from the perspective of its founder and chief proponent.

Author Information

Francis Fallon
Email: Fallonf@stjohns.edu
St. John's University
U. S. A.