Reference: *Biol. Bull.* **215:** 216–242. (December 2008)

Consciousness as Integrated Information: a Provisional Manifesto

GIULIO TONONI

Department of Psychiatry, University of Wisconsin, Madison, Wisconsin

Abstract. The integrated information theory (IIT) starts from phenomenology and makes use of thought experiments to claim that consciousness is integrated information. Specifically: (i) the quantity of consciousness corresponds to the amount of integrated information generated by a complex of elements; (ii) the quality of experience is specified by the set of informational relationships generated within that complex. Integrated information (Φ) is defined as the amount of information generated by a complex of elements, above and beyond the information generated by its parts. Qualia space (Q) is a space where each axis represents a possible state of the complex, each point is a probability distribution of its states, and arrows between points represent the informational relationships among its elements generated by causal mechanisms (connections). Together, the set of informational relationships within a complex constitute a shape in Q that completely and univocally specifies a particular experience. Several observations concerning the neural substrate of consciousness fall naturally into place within the IIT framework. Among them are the association of consciousness with certain neural systems rather than with others; the fact that neural processes underlying consciousness can influence or be influenced by neural processes that remain unconscious; the reduction of consciousness during dreamless sleep and generalized seizures; and the distinct role of different cortical architectures in affecting the quality of experience. Equating consciousness with integrated information carries several implications for our view of nature.

Received 20 August 2008; accepted 10 October 2008.

* To whom correspondence should be addressed. E-mail: gtononi@wisc.edu

Abbreviations: Φ, integrated information; IIT, integrated information theory; MIP, minimum information partition.

INTRODUCTION

Everybody knows what consciousness is: it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream. It is also all we are and all we have: lose consciousness and, as far as you are concerned, your own self and the entire world dissolve into nothingness.

Yet almost everybody thinks that understanding consciousness at the fundamental level is currently beyond the reach of science. The best we can do, it is often argued, is gather more and more facts about the neural correlates of consciousness—those aspects of brain function that change when some aspects of consciousness change—and hope that one day we will come up with an explanation. Others are more pessimistic: we may learn all about the neural correlates of consciousness and still not understand why certain physical processes seem to generate experience while others do not.

It is not that we do not know relevant facts about consciousness. For example, we know that the widespread destruction of the cerebral cortex leaves people permanently unconscious (vegetative), whereas the complete removal of the cerebellum, even richer in neurons, hardly affects consciousness. We also know that neurons in the cerebral cortex remain active throughout sleep, yet at certain times during sleep consciousness fades, while at other times we dream. Finally, we know that different parts of the cortex influence different qualitative aspects of consciousness: damage to certain parts of the cortex can impair the experience of color, whereas other lesions may interfere with the perception of shapes. In fact, increasingly refined neuroscientific tools are uncovering increasingly precise aspects of the neural correlates of consciousness (Koch, 2004). And yet, when it comes to *explaining why* experience blossoms in the cortex and not in the cerebellum, why certain stages of sleep are experientially underprivileged, or why some

cortical areas endow our experience with colors and others with sound, we are still at a loss.

Our lack of understanding is manifested most clearly when scientists are asked questions about consciousness in "difficult" cases. For example, is a person with akinetic mutism—awake with eyes open, but mute, immobile, and nearly unresponsive—conscious or not? How much consciousness is there during sleepwalking or psychomotor seizures? Are newborn babies conscious, and to what extent? Are animals conscious? If so, are some animals more conscious than others? Can they feel pain? Does a bat feel space the same way we do? Can bees experience colors, or merely react to them? Can a conscious artifact be constructed with non-neural ingredients? I believe it is fair to say that no consciousness expert, if there is such a job description, can be confident about the correct answer to such questions. This is a remarkable state of affairs. Just consider comparable questions in physics: Do stars have mass? Do atoms? How many different kinds of atoms and elementary particles are there, and of what are they made? Is energy conserved? And how can it be measured? Or consider biology: What are species, and how do they evolve? How are traits inherited? How do organisms develop? How is energy produced from nutrients? How does echolocation work in bats? How do bees distinguish among colors? And so on. Obviously, we expect satisfactory answers by any competent physicist and biologist.

What's the matter with consciousness, then, and how should we proceed? Early on, I came to the conclusion that a genuine understanding of consciousness is possible only if empirical studies are complemented by a theoretical analysis. Indeed, neurobiological facts constitute both challenging paradoxes and precious clues to the enigma of consciousness. This state of affairs is not unlike the one faced by biologists when, knowing a great deal about similarities and differences between species, fossil remains, and breeding practices, they still lacked a theory of how evolution might occur. What was needed, then as now, were not just more facts, but a theoretical framework that could make sense of them.

In what follows, I discuss the integrated information theory of consciousness (IIT; Tononi, 2004)—an attempt to understand consciousness at the fundamental level. To present the theory, I first consider phenomenological thought experiments indicating that subjective experience has to do with the generation of integrated information. Next, I consider how integrated information can be defined mathematically. I then show how basic facts about consciousness and the brain can be accounted for in terms of integrated information. Finally, I discuss how the quality of consciousness can be captured geometrically by the shape of informational relationships within an abstract space called qualia space. I conclude by examining some implications of the theory concerning the place of experience in our view of the world.

A Phenomenological Analysis: Consciousness as Integrated Information

The *integrated information theory (IIT)* of consciousness claims that, at the fundamental level, consciousness is integrated information, and that its quality is given by the informational relationships generated by a complex of elements (Tononi, 2004). These claims stem from realizing that information and integration are the essential properties of our own experience. This may not be immediately evident, perhaps because, being endowed with consciousness most of the time, we tend to take its gifts for granted. To regain some perspective, it is useful to resort to two thought experiments, one involving a photodiode and the other a digital camera.

Information: the photodiode thought experiment

Consider the following: You are facing a blank screen that is alternately on and off, and you have been instructed to say "light" when the screen turns on and "dark" when it turns off. A photodiode—a simple light-sensitive device—has also been placed in front of the screen. It contains a sensor that responds to light with an increase in current and a detector connected to the sensor that says "light" if the current is above a certain threshold and "dark" otherwise. The first problem of consciousness reduces to this: when you distinguish between the screen being on or off, you have the subjective experience of seeing light or dark. The photodiode can also distinguish between the screen being on or off, but presumably it does not have a subjective experience of light and dark. What is the key difference between you and the photodiode?

According to the IIT, the difference has to do with how much information is generated when that distinction is made. Information is classically defined as reduction of uncertainty: the more numerous the alternatives that are ruled out, the greater the reduction of uncertainty, and thus the greater the information. It is usually measured using the entropy function, which is the logarithm of the number of alternatives (assuming they are equally likely). For example, tossing a fair coin and obtaining heads corresponds to $\log_2(2) = 1$ bit of information, because there are just two alternatives; throwing a fair die yields $\log_2(6) = 2.59$ bits of information, because there are six.

Let us now compare the photodiode with you. When the blank screen turns on, the mechanism in the photodiode tells the detector that the current from the sensor is above rather than below the threshold, so it reports "light." In performing this discrimination between two alternatives, the detector in the photodiode generates $\log_2(2) = 1$ bit of information. When you see the blank screen turn on, on the other hand,

the situation is quite different. Though you may think you are performing the same discrimination between light and dark as the photodiode, you are in fact discriminating among a much larger number of alternatives, thereby generating many more bits of information.

This is easy to see. Just imagine that, instead of turning light and dark, the screen were to turn red, then green, then blue, and then display, one after the other, every frame from every movie that was ever produced. The photodiode, inevitably, would go on signaling whether the amount of light for each frame is above or below its threshold: to a photodiode, things can only be one of two ways, so when it reports "light," it really means just "this way" *versus* "that way." For you, however, a light screen is different not only from a dark screen, but from a multitude of other images, so when you say "light," it really means this specific way *versus* countless other ways, such as a red screen, a green screen, a blue screen, this movie frame, that movie frame, and so on for every movie frame (not to mention for a sound, smell, thought, or any combination of the above). Clearly, each frame looks different to you, implying that some mechanism in your brain must be able to tell it apart from all the others. So when you say "light," whether you think about it or not (and you typically won't), you have just made a discrimination among a very large number of alternatives, and thereby generated many bits of information.

This point is so deceivingly simple that it is useful to elaborate a bit on why, although a photodiode may be as good as we are in detecting light, it cannot possibly see light the way we do—in fact, it cannot possibly "see" anything at all. Hopefully, by realizing what the photodiode lacks, we may appreciate what allows us to consciously "see" the light.

The key is to realize how the many discriminations we can do, and the photodiode cannot, affect the *meaning* of the discrimination at hand, the one between light and dark. For example, the photodiode has no mechanism to discriminate colored from achromatic light, even less to tell which particular color the light might be. As a consequence, all light is the same to it, as long as it exceeds a certain threshold. So for the photodiode, "light" cannot possibly mean achromatic as opposed to colored, not to mention of which particular color. Also, the photodiode has no mechanism to distinguish between a homogeneous light and a bright shape—any bright shape—on a darker background. So for the photodiode, light cannot possibly mean full field as opposed to a shape—any of countless particular shapes. Worse, the photodiode does not even know that it is detecting a visual attribute (the "visualness" of light) as it has no mechanism to tell visual attributes, such as light or dark, from non-visual ones, such as hot and cold, light or heavy, loud or soft, and so on. As far as it knows, the photodiode might just as well be a thermistor—it has no way of knowing whether it is sensing light *versus* dark or hot *versus* cold.

In short, the only specification a photodiode can make is whether things are this or that way: any further specification is impossible because it does not have mechanisms for it. Therefore, when the photodiode detects "light," such "light" cannot possibly mean what it means for us; it does not even mean that it is a visual attribute. By contrast, when we see "light" in full consciousness, we are implicitly being much more specific: we simultaneously specify that things are this way rather than that way (light as opposed to dark), that whatever we are discriminating is not colored (in any particular color), does not have a shape (any particular one), is visual as opposed to auditory or olfactory, sensory as opposed to thought-like, and so on. To us, then, light is much more meaningful precisely because we have mechanisms that can discriminate this particular state of affairs we call "light" against a large number of alternatives.

According to the IIT, it is all this added meaning, provided implicitly by *how* we discriminate pure light from all these alternatives, that increases the level of consciousness. This central point may be appreciated either by "subtraction" or by "addition." By subtraction, one may realize that our being conscious of "light" would degrade more and more—would lose its non-coloredness, its non-shapedness, would even lose its visualness—as its meaning is progressively stripped down to just "one of two ways," as with the photodiode. By addition, one may realize that we can only see "light" as we see it, as progressively more and more meaning is added by specifying how it differs from countless alternatives. Either way, the theory says that the more specifically one's mechanisms discriminate between what pure light is and what it is not (the more they specify what light means), the more one is conscious of it.

Integration: the camera thought experiment

Information—the ability to discriminate among a large number of alternatives—may thus be essential for consciousness. However, information always implies a point of view, and we need to be careful about what that point of view might be. To see why, consider another thought experiment, this time involving a digital camera, say one whose sensor chip is a collection of a million binary photodiodes, each sporting a sensor and a detector. Clearly, taken as a whole, the camera's detectors could distinguish among $2^{1,000,000}$ alternative states, an immense number, corresponding to 1 million bits of information. Indeed, the camera would easily respond differently to every frame from every movie that was ever produced. Yet few would argue that the camera is conscious. What is the key difference between you and the camera?

According to the IIT, the difference has to do with integrated information. From the point of view of an external observer, the camera may be considered as a single system with a repertoire of $2^{1,000,000}$ states. In reality, how-

ever, the chip is not an integrated entity: since its 1 million photodiodes have no way to interact, each photodiode performs its own local discrimination between a low and a high current completely independent of what every other photodiode might be doing. In reality, the chip is just a collection of 1 million independent photodiodes, each with a repertoire of two states. In other words, there is no intrinsic point of view associated with the camera chip as a whole. This is easy to see: if the sensor chip were cut into 1 million pieces each holding its individual photodiode, the performance of the camera would not change at all.

By contrast, you discriminate among a vast repertoire of states as an integrated system, one that cannot be broken down into independent components each with its own separate repertoire. Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and that is experienced from a single point of view. For example, the experience of a red square cannot be decomposed into the separate experience of red and the separate experience of a square. Similarly, experiencing the full visual field cannot be decomposed into experiencing separately the left half and the right half: such a possibility does not even make sense to us, since experience is always whole. Indeed, the only way to split an experience into independent experiences seems to be to split the brain in two, as in patients who underwent the section of the corpus callosum to treat severe epilepsy (Gazzaniga, 2005). Such patients do indeed experience the left half of the visual field independently of the right side, but then the surgery has created two separate consciousnesses instead of one. Mechanistically then, underlying the unity of experience must be causal interactions among certain elements within the brain. This means that these elements work together as an integrated system, which is why their performance, unlike that of the camera, breaks down if they are disconnected.

A Mathematical Analysis: Quantifying Integrated Information

This phenomenological analysis suggests that, to generate consciousness, a physical system must be able to discriminate among a large repertoire of states (information) *and* it must be unified; that is, it should be doing so as a single system, one that is not decomposable into a collection of causally independent parts (integration). But how can one measure integrated information? As I explain below, the central idea is to quantify the information generated by a system, above and beyond the information generated independently by its parts (Tononi, 2001, 2004; Balduzzi and Tononi, 2008).[1]

Information

First, we must evaluate how much information is generated by the system. Consider the system of two binary units in Figure 1, which can be thought of as an idealized version of a photodiode composed of a sensor S and a detector D. The system is characterized by a state it is in, which in this case is 11 (first digit for the sensor, second digit for the detector), and by a mechanism. This is mediated by a connection (arrow) between the sensor and the detector that implements a causal interaction: in this case, the elementary mechanism of the system is that the detector checks the state of the sensor and turns on if the sensor is on, and off otherwise (more generally, the specific causal interaction can be described by an input-output table).

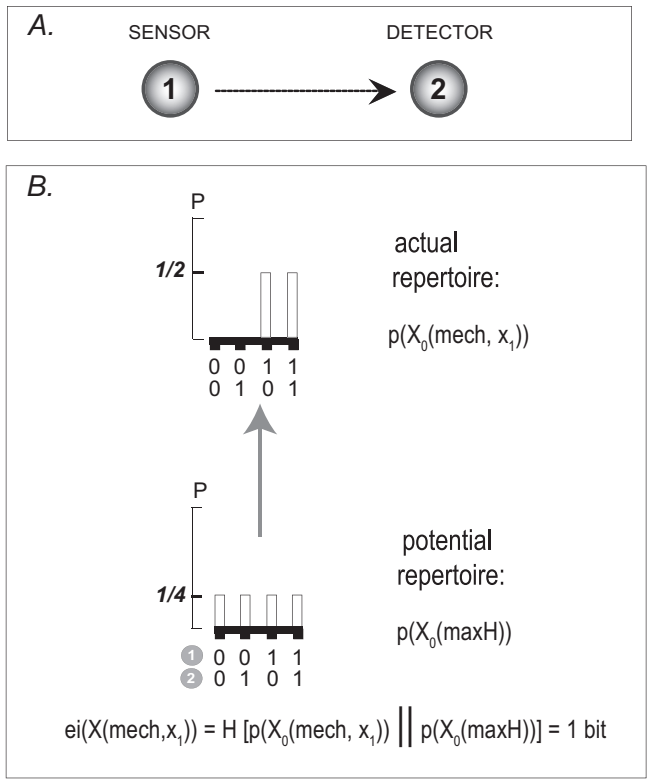

Figure 1. Effective information. (A) A "photodiode" consisting of a sensor and detector unit. The photodiode's mechanism is such that the detector unit turns on if the sensor's current is above a threshold. Here both units are on (binary 1, indicated in gray). (B) For the entire system (sensor unit, detector unit) there are four possible states: (00,01,10,11). The potential distribution $p(X_0(\text{maxH})) = (1/4,1/4,1/4,1/4)$ is the maximum entropy distribution on the four states. Given the photodiode's mechanism and the fact that the detector is on, the sensor must have been on. Thus, the photodiode's mechanism and its current state specifies the following distribution: two of the four possible states (00,01) are ruled out; the other two states (10,11) are equally likely since they are indistinguishable to the mechanism (the prior state of the detector makes no difference to the current state of the sensor). The actual distribution is therefore $p(X_0(\text{mech}, x_1)) = (0,0,1/2,1/2)$. Relative entropy (Kullback-Leibler divergence) between two probability distributions p and q is $H[p|q] = p_i \log_2 p_i/q_i$, so the effective information $ei(X(\text{mech}, x_1))$ associated with output $x_1 = 11$ is 1 bit (effective information is the entropy of the actual relative to the potential distributions).

Potentially, a system of two binary elements could be in any of four possible states (00,01,10,11) with equal proba-

bility: p = (1/4,1/4,1/4,1/4). Formally, this *potential* (*a priori*) repertoire is represented by the maximum entropy or uniform distribution of possible system states at time t=0, which expresses complete uncertainty ($p(X_0(maxH))$). Considering the potential repertoire as the set of all possible input states, the particular mechanism X(mech) of this system can be thought of as specifying a *forward* repertoire—the probability distribution of output states produced by the system when perturbed with all possible input states. But the system is actually in a particular output state (in this case, at time t=1, x_1 = 11). In actuality, a system with this mechanism being in state 11 *specifies* that the previous system state x_0 must have been either 11 or 10, rather than 00 or 01, corresponding to p = (0,0,1/2,1/2) (in this system, there is no mechanism to specify the detector state, which remains uncertain). Formally, then, the mechanism and the state 11 specify an *actual* (*a posteriori*) distribution or repertoire of system states $p(X_0(mech,x_1))$ at time t=0 that could have caused (led to) x_1 at time t=1, while ruling out (giving probability zero to) states that could not. In this way, the system's mechanism and state constitute information (about the system's previous state), in the classic sense of reduction of uncertainty or ignorance. More precisely, the system's mechanism and state generate 1 bit of information by distinguishing between things being one way (11 or 10, which remain indistinguishable to it) rather than another way (00 or 01, which also remain indistinguishable to it).

In general, the information generated when a system characterized by a certain mechanism in a particular state can be measured by the *relative entropy* H between the actual and the potential repertoires ("relative to" is indicated by ‖), captured by the *effective information* (ei):

$$ei(X(mech,x_1) = H[\,p(X_0(mech,x_1))\|p(X_0(maxH))\,]$$

Relative entropy, also known as Kullback-Leibler divergence, is a difference between probability distributions (Cover and Thomas, 2006): if the distributions are identical, relative entropy is zero; the more different they are, the higher the relative entropy.[2] Figuratively, the system's mechanism and state generate information by sharpening the uniform distribution into a less uniform one—this is how much uncertainty is reduced. Clearly, the amount of effective information generated by a system is high if it has a large potential repertoire and a small actual repertoire, since a large number of initial states are ruled out. By contrast, the information generated is little if the system's repertoire is small, or if many states could lead to the current outcome, since few states are ruled out. For instance, if noise dominates (any state could have led to the current one), no alternatives are ruled out, and no information is generated.

Since effective information is implicitly specified once a mechanism and state are specified, it can be considered to be an "intrinsic" property of a system. To calculate it explicitly, from an extrinsic perspective, one can perturb the system in all possible ways (*i.e.,* try out all possible input states, corresponding to the maximum entropy distribution or potential repertoire) to obtain the forward repertoire of output states given the system's mechanism. Finally one can calculate, using Bayes' rule, the actual repertoire given the system's state (Balduzzi and Tononi, 2008).[3]

Integration

Second, we must find out how much of the information generated by a system is integrated information; that is, how much information is generated by a single entity, as opposed to a collection of independent parts. The idea here is to consider the parts of the system independently, ask how much information they generate by themselves, and compare it with the information generated by the system as a whole.

This can be done by resorting again to relative entropy to measure the difference between the probability distribution generated by the system as a whole ($p(X_0(mech,x_1))$, the actual repertoire of the system x) with the probability distribution generated by the parts considered independently ($\Pi p({}^kM_0(mech,\mu_1))$, the product of the actual repertoire of the parts ${}^kM$). Integrated information is indicated with the symbol Φ (the vertical bar "I" stands for information, the circle "O" for integration):

$$\Phi\ (X(mech,x_1))$$
$$= H[\,p(X_0(\mathrm{mech},x_1))\|\textstyle\prod p({}^kM_0(\mathrm{mech},\mu_1))\,]\ for\ {}^kM_0 \in MIP$$

That is, the actual repertoire for each part is specified by causal interactions internal to each part, considered as a system in its own right, while external inputs are treated as a source of extrinsic noise. The comparison is made with the particular decomposition of the system into parts that leaves the least information unaccounted for. This *minimum information partition* (MIP) decomposes the system into its *minimal parts*.

To see how this works, consider two of the million photodiodes in the digital camera (Fig. 2, left). By turning on or off depending on its input, each photodiode generates 1 bit of information, just as we saw before. Considered independently, then, two photodiodes generate 2 bits of information, and 1 million photodiodes generate 1 million bits of information. However, as shown in the figure, the product of the actual distributions generated independently by the parts is identical to the actual distribution for the system. Therefore, the relative entropy between the two distributions is zero: the system generates no integrated information ($\Phi\ (X(mech,x_1)) = 0$) above and beyond what is generated by its parts.

Clearly, for integrated information to be high, a system must be connected in such a way that information is gen-

erated by causal interactions *among* rather than *within* its parts. Thus, a system can generate integrated information only to the extent that it cannot be decomposed into informationally independent parts. A simple example of such a system is shown in Figure 2 (right). In this case, the interaction between the minimal parts of the system generates information above and beyond what is accounted for by the parts by themselves (Φ ($X(mech,x_1)$) > 0).

In short, integrated information captures the information generated by causal interactions in the whole, over and above the information generated by the parts.[4]

Complexes

Finally, by measuring Φ values for all subsets of elements within a system, we can determine which subsets form *complexes*. Specifically, a complex X is a set of elements that generate integrated information ($\Phi > 0$) that is not fully contained in some larger set of higher Φ (Fig. 3). A complex, then, can be properly considered to form a single entity having its own, intrinsic "point of view" (as opposed to being treated as a single entity from an outside, extrinsic point of view). Since integrated information is generated *within* a complex and not outside its boundaries, experience is necessarily private and related to a single point of view or perspective (Tononi and Edelman, 1998; Tononi, 2004). A given physical system, such as a brain, is likely to contain more than one complex, many small ones with low Φ values, and perhaps a few large ones (Tononi and Edelman, 1998; Tononi, 2004). In fact, at any given time there may be a single *main complex* of comparatively much higher Φ that underlies the dominant experience (a main complex is such that its subsets have strictly lower Φ). As shown in Figure 3, a main complex can be embedded into larger complexes of lower Φ. Thus, a complex can be casually connected, through *ports-in* and *ports-out*, to elements that are not part of it. According to the IIT, such elements can indirectly influence the state of the main complex without contributing directly to the conscious experience it generates (Tononi and Sporns, 2003).

A Neurobiological Reality Check: Accounting for Empirical Observations

Can this approach account, at least in principle, for some of the basic facts about consciousness that have emerged from decades of clinical and neurobiological observations? Measuring Φ and finding complexes is not easy for realistic systems, but it can be done for simple networks that bear some structural resemblance to different parts of the brain (Tononi, 2004; Balduzzi and Tononi, 2008).

For example, by using computer simulations, it is possible to show that high Φ requires networks that conjoin functional specialization (due to its specialized connectivity; each element has a unique functional role within the network) with functional integration (there are many pathways for interactions among the elements, Fig. 4A.). In very rough terms, this kind of architecture is characteristic of the mammalian corticothalamic system: different parts of the cerebral cortex are specialized for different functions, yet a vast network of connections allows these parts to interact profusely. And indeed, as much neurological evidence indicates (Posner and Plum, 2007), the corticothalamic system is precisely the part of the brain that cannot be severely impaired without loss of consciousness.

Conversely, Φ is low for systems that are made up of small, quasi-independent modules (Fig. 4B; Tononi, 2004; Balduzzi and Tononi, 2008). This may be why the cerebellum, despite its large number of neurons, does not contribute much to consciousness: its synaptic organization is such that individual patches of cerebellar cortex tend to be activated independently of one another, with little interaction between distant patches (Bower, 2002).

Computer simulations also show that units along multiple, segregated incoming or outgoing pathways are not incorporated within the repertoire of the main complex (Fig. 4C; Tononi, 2004; Balduzzi and Tononi, 2008). This may be why neural activity in afferent pathways (perhaps as far as V1), though crucial for triggering this or that conscious experience, does not contribute directly to conscious experience; nor does activity in efferent pathways (perhaps starting with primary motor cortex), though it is crucial for reporting each different experience.

The addition of many parallel cycles also generally does not change the composition of the main complex, although Φ values can be altered (Fig. 4D). Instead, cortical and subcortical cycles or loops implement specialized subroutines that are capable of influencing the states of the main corticothalamic complex without joining it. Such informationally insulated cortico-subcortical loops could constitute the neural substrates for many unconscious processes that can affect and be affected by conscious experience (Baars, 1988; Tononi, 2004), such as those that enable object recognition, language parsing, or translating our vague intentions into the right words.

At this stage, it is hard to say precisely which cortical circuits may work as a large complex of high Φ, and which instead may remain informationally insulated. Does the dense mesial connectivity revealed by diffusion spectral imaging (Hagmann *et al.,* 2008) constitute the "backbone" of a corticothalamic main complex? Do parallel loops through basal ganglia implement informationally insulated subroutines? Are primary sensory cortices organized like massive afferent pathways to a main complex higher up in the cortical hierarchy (Koch, 2004)? Is much of prefrontal cortex organized like a massive efferent pathway? Do certain cortical areas, such as those belonging to the dorsal visual stream, remain partly segregated from the main complex? Unfortunately, answering these questions and prop-

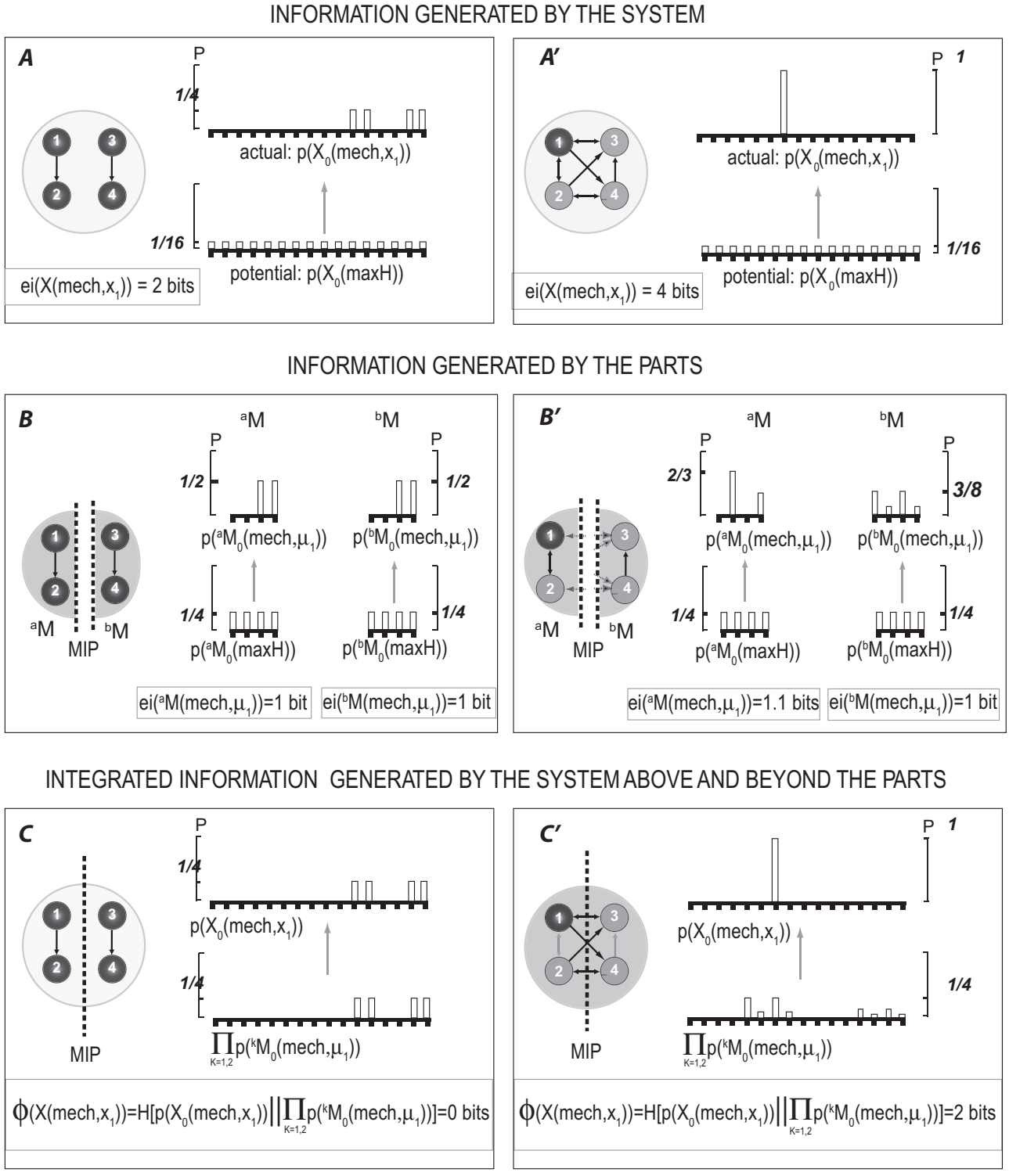

Figure 2. Integrated information. *Left-hand side:* two photodiodes in a digital camera. (A) Information generated by the system as a whole. The system as a whole generates 2 bits of effective information by specifying that n_1 and n_3 must have been on. (B) Information generated by the parts. The minimum information partition (MIP) is the decomposition of a system into (minimal) parts, that is, the decomposition that leaves the least information unaccounted for. Here the parts are two photodiodes. (C) The information generated by the system as a whole is completely accounted for by the information generated by its parts. In this case, the actual repertoire of the whole is identical to the combined actual repertoires of the parts (the product of their

erly testing the predictions of the theory requires a much better understanding of cortical neuroanatomy than is currently available.

Other simulations show that the effects of cortical disconnections are readily captured in terms of integrated information (Tononi, 2004): a "callosal" cut produces, out of a large complex corresponding to the connected corticothalamic system, two separate complexes, in line with many studies of split-brain patients (Gazzaniga, 2005). However, because there is great redundancy between the two hemispheres, their Φ value is not greatly reduced compared to when they form a single complex. Functional disconnections may also lead to a restriction of the neural substrate of consciousness, as is seen in neurological neglect phenomena, in psychiatric conversion and dissociative disorders, and possibly during dreaming and hypnosis. It is also likely that certain attentional phenomena may correspond to changes in the composition of the main complex underlying consciousness (Koch and Tsuchiya, 2007). The attentional blink,[5] where a fixed sensory input may at times make it to consciousness and at times not, may also be due to changes in functional connectivity: access to the main corticothalamic complex may be enabled or not based on dynamics intrinsic to the complex (Dehaene *et al.,* 2003). Similarly, binocular rivalry[6] may be related, at least in part, to dynamic changes in the composition of the main corticothalamic complex caused by transient changes in functional connectivity. Computer simulations confirm that functional disconnection can reduce the size of a complex and reduce its capacity to integrate information (Tononi, 2004). While it is not easy to determine, at present, whether a particular group of neurons is excluded from the main complex because of hard-wired anatomical constraints or is transiently disconnected due to functional changes, the set of elements underlying consciousness is not static, but form a "*dynamic complex*" or "*dynamic core*" (Tononi and Edelman, 1998).

Computer simulations also indicate that the capacity to integrate information is reduced if neural activity is extremely high and near-synchronous, due to a dramatic decrease in the repertoire of discriminable states (Fig. 4E; Balduzzi and Tononi, 2008). This reduction in degrees of freedom could be the reason that consciousness is reduced or eliminated in absence seizure (petit mal) and other conditions during which neural activity is both high and synchronous (Blumenfeld and Taylor, 2003).

The most common example of a marked change in the level of experience is the fading of consciousness that occurs during certain periods of sleep. Subjects awakened in deep NREM (non–rapid eye movement) sleep, especially early in the night, often report that they were not aware of themselves or of anything else, though cortical and thalamic neurons remain active. Awakened at other times, mainly during REM sleep or during lighter periods of NREM sleep later in the night, they report dreams characterized by vivid images (Hobson *et al.,* 2000). From the perspective of integrated information, a reduction of consciousness during early sleep would be consistent with the bistability of cortical circuits during deep NREM sleep. Due to changes in intrinsic and synaptic conductances triggered by neuromodulatory changes (*e.g.,* low acetylcholine), cortical neurons cannot sustain firing for more than a few hundred milliseconds and invariably enter a hyperpolarized down-state. Shortly afterward, they inevitably return to a depolarized up-state (Steriade *et al.,* 2001). Indeed, computer simulations show that values of Φ are low in systems with such bistable dynamics (Fig. 4F, Balduzzi and Tononi, 2008). Consistent with these observations, studies using TMS, a technique for stimulating the brain non-invasively, in conjunction with high-density EEG, show that early NREM sleep is associated either with a breakdown of the effective connectivity among cortical areas, and thereby with a loss of integration (Massimini *et al.,* 2005, 2007), or with a stereotypical global response suggestive of a loss of repertoire and thus of information (Massimini *et al.,* 2007). Similar changes are seen in animal studies of anesthesia (Alkire *et al.,* 2008).

Finally, consciousness not only requires a neural substrate with appropriate anatomical structure and appropriate physiological parameters, it also needs time (Bachmann, 2000). The theory predicts that the time requirement for the generation of conscious experience in the brain emerges directly from the time requirements for the build-up of an integrated repertoire among the elements of the corticothalamic main complex so that discriminations can be highly informative (Tononi, 2004; Balduzzi and Tononi, unpubl.). To give an obvious example, if one were to perturb half of the elements of the main complex for less than a millisecond, no perturbations would produce any effect on the other half within this time window, and Φ would be zero. After, say, 100 ms, however, there is enough time for differential effects to be manifested, and Φ should grow.

---

respective probability distributions), so that relative entropy is zero. The system generates no information above and beyond the parts, so it cannot be considered a single entity. *Right-hand side:* an integrated system. Elements in the system are on if they receive two or more spikes. The system is in state $x_1 = 1000$. (A′) The mechanism specifies a unique prior state that can cause state x_1, so the system generates 4 bits of effective information. All other initial states are ruled out, since they cause different outputs. (B′) Effective information generated by the two minimal parts, considered as systems in their own right. External inputs are treated as extrinsic noise. (C′) Integrated information is information generated by the whole (black arrows) over and above the parts (gray arrows). In this case, the actual repertoire of the whole is different from the combined actual repertoires of the parts, and the relative entropy is 2 bits. The system generates information above and beyond the parts, so it can be considered a single entity (a complex).

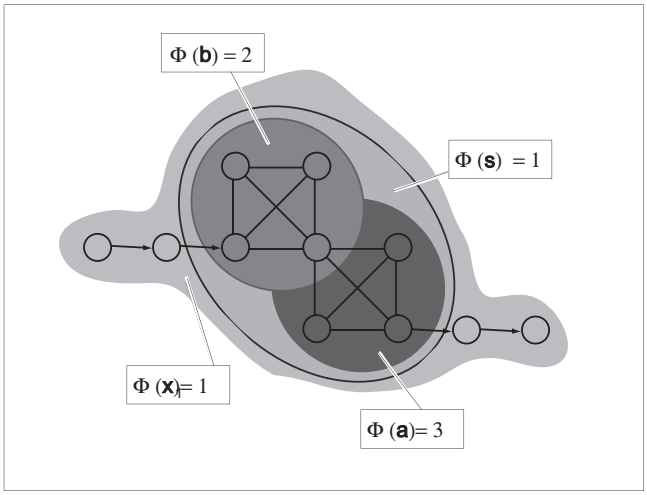

Figure 3. Complexes. In this system, the mechanism is that elements fire in response to an odd number of spikes on their afferent connections (links without arrows are bidirectional connections). Analyzing the system in terms of integrated information shows that the system constitutes a complex (x, light gray) that contains three smaller complexes (s,a,b, in different shades of gray). Observe that *(i)* complexes can overlap; *(ii)* a complex can interact causally with elements not part of it; *(iii)* groups of elements with identical architectures (a and b) generate different amounts of integrated information, depending on their ports-in and ports-out.

The Quality of Consciousness: Characterizing Informational Relationships

If the amount of integrated information generated by different brain structures (or by the same structure functioning in different ways) can in principle account for changes in the level of consciousness, what is responsible for the quality of each particular experience? What determines that colors look the way they do and are different from the way music sounds? Once again, empirical evidence indicates that different qualities of consciousness must be contributed by different cortical areas. Thus, damage to certain parts of the cerebral cortex forever eliminates our ability to experience color (whether perceived, imagined, remembered, or dreamt), whereas damage to other parts selectively eliminates our ability to experience visual shapes. There is obviously something about different parts of the cortex that can account for their different contribution to the quality of experience. What is this something?

The IIT claims that, just as the *quantity* of consciousness generated by a complex of elements is determined by the amount of integrated information it generates above and beyond its parts, the *quality* of consciousness is determined by the set of all the informational relationships its mechanisms generate. That is, *how* integrated information is generated within a complex determines not only the amount of consciousness it has, but also what kind of consciousness.

Consider again the photodiode thought experiment. As I discussed before, when the photodiode reacts to light, it can only tell that things are one way rather than another way. On the other hand, when we see "light," we discriminate against many more states of affairs, and thus generate much more information. In fact, I argued that "light" means what it means and becomes conscious "light" *by virtue of* being not just the opposite of dark, but also different from any color, any shape, any combination of colors and shapes, any frame of every possible movie, any sound, smell, thought, and so on.

What needs to be emphasized at this point is that discriminating "light" against all these alternatives implies not just picking one thing out of "everything else" (an undifferentiated bunch), but distinguishing at once, in a specific way, between each and every alternative. Consider a very simple example: a binary counter capable of discriminating among the four numbers: 00, 01, 10, 11. When the counter says binary "3," it is not just discriminating 11 from everything else as an undifferentiated bunch, otherwise it would not be a counter, but a 11 detector. To be a counter, the system must be able to tell 11 apart from 00 as well as from 10 as well as from 01 in different, specific ways. It does so, of course, by making choices through its mechanisms; for example: is this the first or the second digit? Is it a 0 or a 1? Each mechanism adds its specific contribution to the discrimination they perform together. Similarly, when we see light, mechanisms in our brain are not just specifying "light" with respect to a bunch of undifferentiated alternatives. Rather, these mechanisms are specifying that light is what it is by virtue of being different, in this and that specific way, from every other alternative—from dark to any color, to any shape, movie frame, sound or smell, and so on.

In short, generating a large amount of integrated information entails having a highly structured set of mechanisms that allow us to make many nested discriminations (choices) as a single entity. According to the IIT, these mechanisms working together generate integrated information by specifying a set of informational relationships that completely and univocally determine the quality of experience.

Experience as a shape in qualia space

To see how this intuition can be given a mathematical formulation, let us consider again a complex of n binary elements $X(mech,x_1)$ having a particular mechanism and being in a particular state. The mechanism of the system is implemented by a set of connections X^{conn} among its elements. Let us now suppose that each possible state of the system constitutes an axis or dimension of a *qualia space* (Q) having 2^n dimensions. Each axis is labeled with the probability p for that state, going from 0 to 1, so that a repertoire (*i.e.,* a probability distribution on the possible states of the complex) corresponds to a point in Q (Fig. 5).

Let us now examine how the connections among the elements of the complex specify probability distributions; that is, how a set of mechanisms specifies a set of informa-

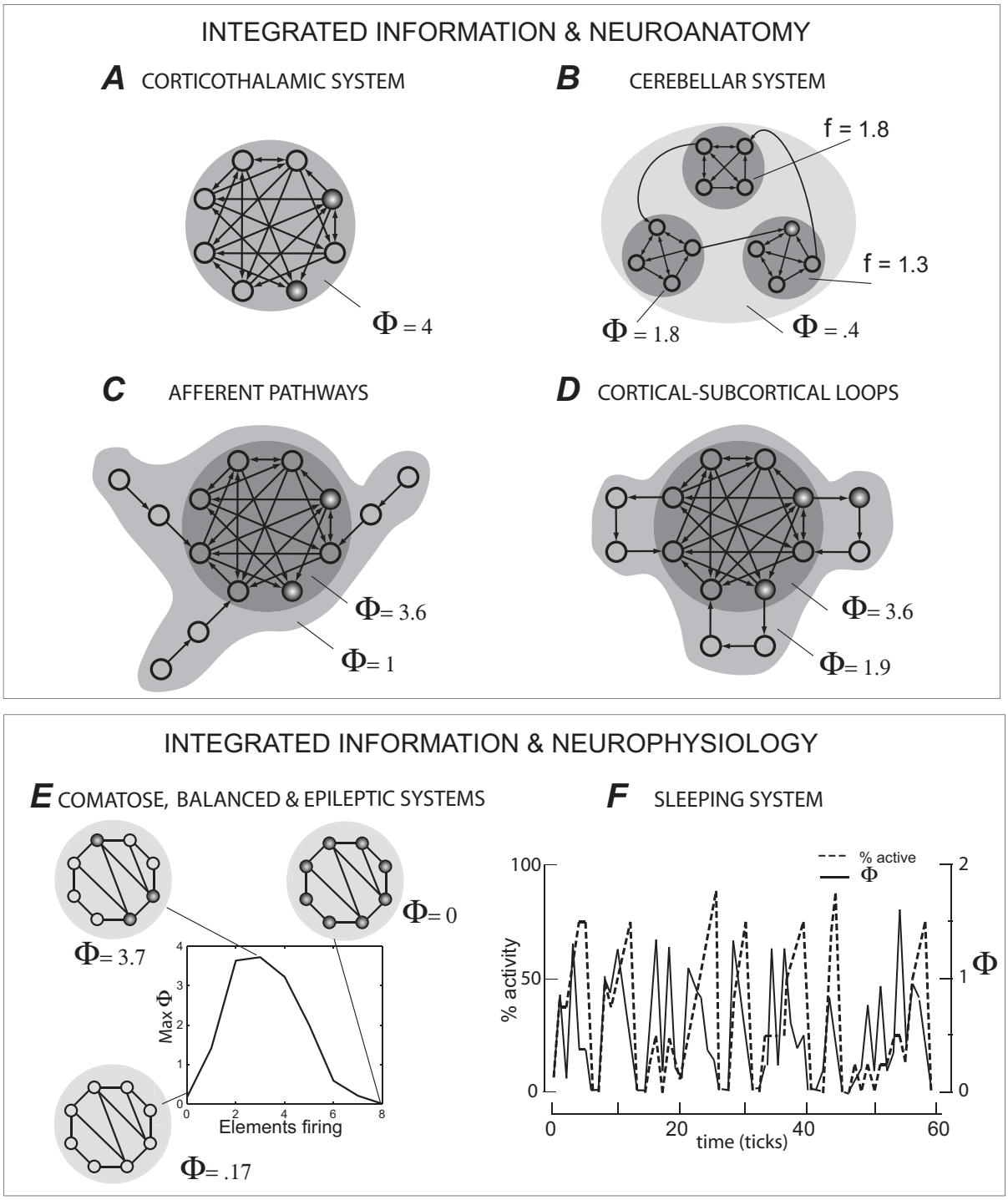

Figure 4. Relating integrated information to neuroanatomy and neurophysiology. Elements fire in response to two or more spikes (except elements targeted by a single connection, which copy their input); links without arrows are bidirectional. (A) Computing Φ in simple models of neuroanatomy suggests that a functionally integrated and functionally specialized network—like the corticothalamic system—is well suited to generating high values of Φ. (B, C, D) Architectures modeled on the cerebellum, afferent pathways, and cortical-subcortical loops give rise to complexes containing more elements, but with reduced Φ compared to the main corticothalamic complex. (E) Φ peaks in balanced states; if too many or too few elements are active, Φ collapses. (F) In a bistable ("sleeping") system (same as in (E)), Φ collapses when the number of firing elements (dotted line) is too high (high % activity), remains low during the "DOWN" state (zero % activity), and only recovers at the onset of the next "UP" state.

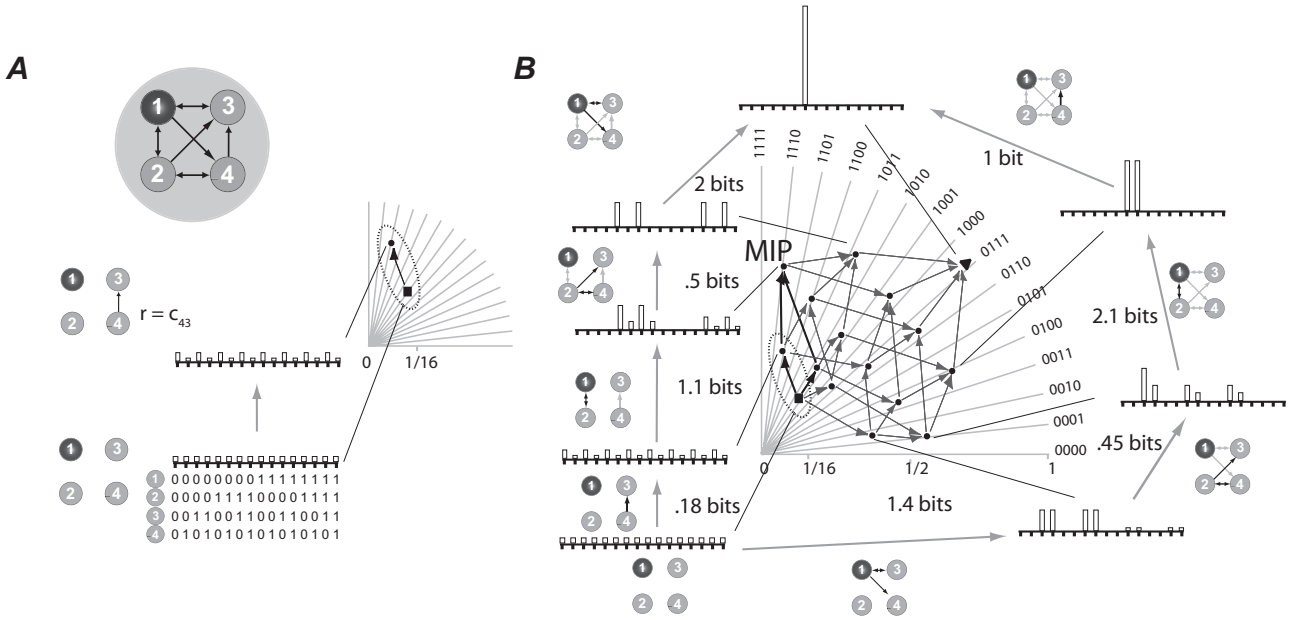

Figure 5. Qualia. (A) The system in the inset is the same as in Fig. 2A′. Qualia (Q)-space for a system of four units is 16-dimensional (one axis per possible state; since axes are displayed flattened onto the page, and points and arrows cannot be properly drawn in 2-dimensions, their position and direction is for illustration only). In state $x_1 = 1000$, the complex generates a quale or shape in Q, as follows. The maximum entropy distribution (the "bottom" of the quale, indicated by a black square) is a point assigning equal probability ($p = 1/16 = 0.0625$) to all 16 system states, close to the origin of the 16-dimensional space. Engaging a single connection "r" between elements 4 and 3 (c_{43}) specifies that, since element n_3 has not fired, the probability of element n_4 having fired in the previous time step is reduced to $p = 0.25$ compared to its maximum entropy value ($p = 0.5$), while the probability of n_4 not having fired is increased to $p = 0.75$. The actual probability distribution of the 16 system states is modified accordingly. Thus, the connection r "sharpens" the maximum entropy distribution into an actual distribution, which is another point in Q. The *q-arrow* linking the two distributions geometrically realizes the *informational relationship* specified by the connection. The length (divergence) of the q-arrow expresses *how much* the connection specifies the distribution (the effective information it generates or relative entropy between the two distributions); the direction in Q expresses *the particular way* in which the connection specifies the distribution. (B) Engaging more connections further sharpens the actual repertoire, specifying new points in Q and the corresponding q-arrows. The figure shows 16 out of the 399 points in the quale, generated by combinations of the four sets of connections. The probability distributions depicted around the quale are representative of the repertoires generated by two q-edges formed by q-arrows that engage the four sets of connections in two different orders (the two representative q-edges start at bottom left—one goes clockwise, the other counter-clockwise; black connections represent those whose contribution is being evaluated; gray connections those whose contribution has already been considered and which provides the context on top of which the q-arrow generated by a black connection begins). Repertoires corresponding to certain points of the quale are shown alongside, as in previous figures. Effective information values (in bits) of the q-arrows in the two q-edges are shown alongside. Together, the q-edges enclose a shape, the quale, which completely specifies the quality of the experience.

tional relationships. First, consider the complex with all connections among its elements disengaged, thus discounting any causal interactions (Fig. 5A). In the absence of a mechanism, the state x_1 provides no information about the system's previous state: from the perspective of a system without causal interactions, all previous states are equally likely, corresponding to the maximum entropy or uniform distribution (the potential repertoire). In Q, this probability distribution is a point projecting onto all axes at $p = 1/2^n$ (probabilities must sum to 1).

Next, consider engaging a single connection (Fig. 5A, the other connections are treated as extrinsic noise). As with the photodiode, the mechanism implemented by that connection and the state the system is in rule out states that could not have caused x_1 and increases the actual probability of states that could have caused x_1, yielding an actual repertoire. In Q, the actual repertoire specified by this connection corresponds to a point projecting onto higher p values on some axes and onto lower p values (or zero) on other axes. Thus, the connection shapes the uniform distribution into a more specific distribution, and thereby generates information (reduces uncertainty). More generally, we can say that the connection specifies an *informational relationship*, that is, a relationship between two probability distributions. This in-

formational relationship can be represented as an arrow in Q (*q-arrow*) that goes from the point corresponding to the maximum entropy distribution ($p = 1/2^{n)}$ to the point corresponding to the actual repertoire specified by that connection. The length (divergence) of the q-arrow expresses *how much* the connection specifies the distribution (the effective information it generates, *i.e.,* the relative entropy between the two distributions); the direction in Q expresses *the particular way* in which the connection specifies the distribution, *i.e.,* a change in position in Q. Similarly, if one considers all other connections taken in isolation, each will specify another q-arrow of a certain length, pointing in a different direction.

Next, consider all possible combinations of connections (Fig. 5B). For instance, consider adding the contribution of the second connection to that of the first. Together, the first and second connections specify another actual *repertoire*—another point in Q-space—and thereby generate more information than either connection alone as they shape the uniform distribution into a more specific distribution. To the tip of the q-arrow specified by the first connection, one can now add a q-arrow bent in the direction contributed by the second connection, forming an "edge" of two q-arrows in Q-space (the same final point is reached by adding the q-arrow due to the first connection on top of the q-arrow specified by the second one). Each combination of connection therefore specifies a *q-edge* made of concatenated q-arrows (component q-arrows). In general, the more connections one considers together, the more the actual repertoire will take shape and differ from the uniform (potential) distribution.

Finally, consider the joint contribution of all connections of the complex (Fig. 5B). As was discussed above, all connections together specify the actual repertoire of the whole. This is the point where all q-edges converge. Together, these q-edges in Q delimit a *quale*, that is, a *shape* in Q, a kind of 2^n-dimensional solid (technically, in more than three dimensions, the "body" of a polytope). The bottom of the quale is the maximum entropy distribution, its edges are q-edges made of concatenated q-arrows, and its top is the actual repertoire of the complex as a whole. The shape of this solid (polytope) is specified by all informational relationships that are generated within the complex by the interactions among its elements (the *effective information matrix*; Tononi, 2004).[7] Note that the same complex of elements, endowed with the same mechanism, will typically generate a different quale or shape in Q depending on the particular state it is in.

It is worth considering briefly a few relevant properties of informational relationships or q-arrows. First, informational relationships are context-dependent (Fig. 6), in the following sense. A *context* can be any point in Q corresponding to the actual repertoire generated by a particular subset of connections. It can be shown that the q-arrow generated by considering the effects of an additional connection (how it further sharpens the actual repertoire) can change in both magnitude and direction depending on the context in which it is considered. In Figure 6, when considered in isolation (null context), the connection "r" between elements 4 and 3 generates a short q-arrow (0.18 bits) pointing in a certain direction. When considered in the full context provided by all other connections (not-r or ¬r), the same connection "r" generates a longer q-arrow (1 bit) pointing in a different direction.

Another property is how removing or adding a set of connections folds or unfolds a quale. The portion of the quale that is generated by a set of connections r (acting in all contexts) is called a *q-fold*. If we remove connection r from the system, all the q-arrows generated by that connection, in all possible contexts, vanish, so the shape of the quale "folds" along the q-fold specified by that connection. Conversely, when the connection is added to a system, the shape of the quale unfolds.

Another important property of q-arrows is *entanglement* (γ, Balduzzi and Tononi, unpubl.). A q-arrow is entangled ($\gamma > 0$) if the underlying connections considered together generate information above and beyond the information they generate separately (note the analogy with Φ). Thus, entanglement characterizes informational relationships (q-arrows) that are more than the sum of their component relationships (component q-arrows, Fig. 6B), just like Φ characterizes systems that are more than the sum of their parts. Geometrically, entanglement "warps" the shape of the quale away from a simple hypercube (where q-arrows are orthogonal to each other). Entanglement has several relevant consequences (Balduzzi and Tononi, unpubl.). For example, an entangled q-arrow can be said to specify a *concept*, in that it groups together certain states of affairs in a way that cannot be decomposed into the mere sum of simpler groupings (see also Feldman, 2003). Moreover, just as Φ can be used to identify complexes, entanglement γ can be used to identify *modes*. By analogy with complexes, *modes* are sets of q-arrows that are more densely entangled than surrounding q-arrows: they can be considered as clusters of informational relationships constituting distinctive "sub-shapes" in Q (see Fig. 8). By analogy with a main complex, an *elementary mode* is such that its component q-arrows have strictly lower γ. As will be briefly discussed below, modes play an important role in understanding the structure of experience.

Some properties of qualia space

What is the relevance of these constructs to understanding the quality of consciousness? It is not easy to become familiar with a complicated multidimensional space nearly impossible to draw, so it may be useful to resort to some metaphors. I have argued that the set of informational rela-

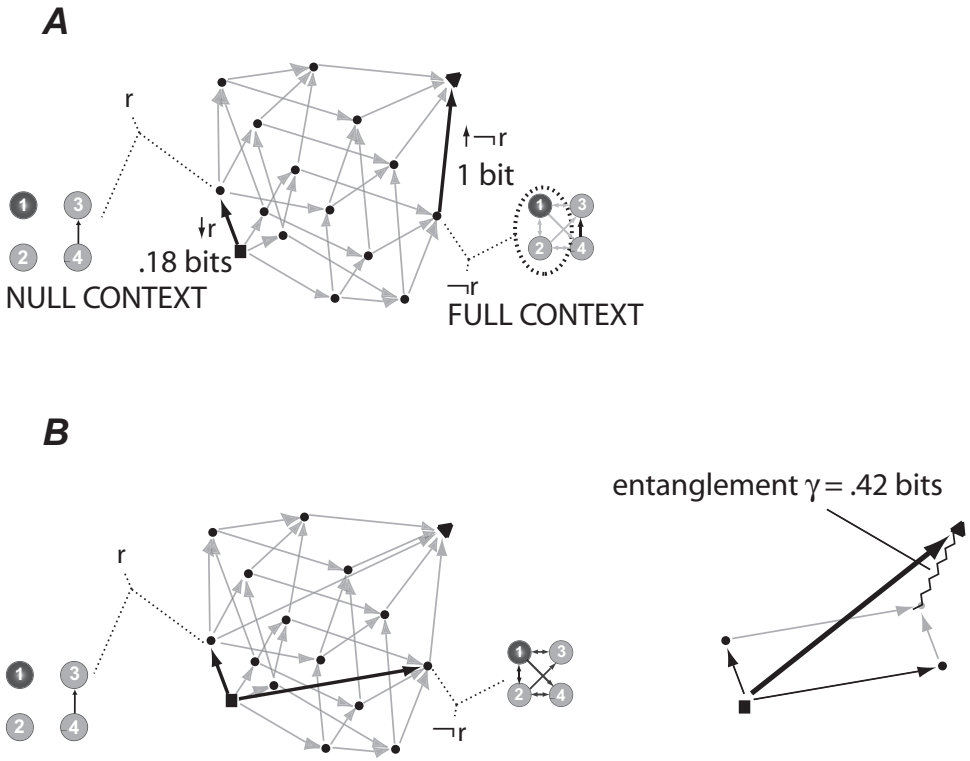

Figure 6. Context and entanglement. (A) Context. The same connection (black arrow between elements 3 and 4) considered in two contexts. At the bottom of the quale (null context, corresponding to the maximum entropy distribution when no other connections are engaged), the connection r generates a q-arrow (called down-set of r, or ↓ r) corresponding to 0.18 bits of information pointing up-left in Q. Near the top of the quale (full context, corresponding to the actual distribution specified by all other connections except for r, indicated as ¬r), r generates a q-arrow (called up-set of non-red, or ↑ ¬r) corresponding to 1 bit of information pointing up-right in Q. (B) Entanglement. Left: the q-arrow generated by the connection "r" and the q-arrow generated by the complementary connections "¬ r" at the bottom of the quale (null context). Right: The product of the two q-arrows (corresponding to independence between the informational relationships specified by the two sets of connections) would be a point corresponding to the vertex of the dotted parallelogram opposite to the bottom. However, "r" and "¬r" jointly specify the actual distribution corresponding to the top of the quale (black triangle). The distance between the probability distribution in Q specified jointly by two sets of connections and their product distribution (zigzag arrow) is the entanglement between the two corresponding q-arrows (how much the composite q-arrow specifies above and beyond its component q-arrows).

tionships in Q generated by the mechanisms of a complex in a given state (q-arrows between repertoires) specify a shape in Q (a quale). Perhaps the most important notion emerging from this approach is that *an experience is a shape in Q*. According to the IIT, *this shape completely and univocally*[8] *specifies the quality of experience.*

It follows that different experiences are, literally, different shapes in Q. For example, when the same system is in a different state (firing pattern), it will typically generate a different shape or quale (even for the same value of Φ). Importantly, if an element turns on, it generates information and meaning not by signifying something (say "red"), which in isolation it cannot, but by changing the shape of the quale. Moreover, experiences are similar if their shape is similar, and different to the extent that their shapes are different. This means that phenomenological similarities and differences can in principle be quantified as similarities and differences between shapes. The set of all shapes generated by the same system in different states provides a geometrical depiction of all its possible experiences.[9]

Note that a quale can only be specified by a mechanism and a particular state—it does not make sense to ask about the quale generated by a mechanism in isolation, or by a state (firing pattern) in isolation. A consequence is that two different systems in the same state can generate two different experiences (*i.e.,* two different shapes). As an extreme example, a system that was to copy one by one the state of the neurons in a human brain, but had no internal connections of its own, would generate no consciousness and no quale (Tononi, 2004; Balduzzi and Tononi, 2008).

By the same token, it is possible that two different systems generate the same experience (*i.e.,* the same shape).

For example, consider again the photodiode, whose mechanism determines that if the current in the sensor exceeds a threshold, the detector turns on. This simple causal interaction is all there is, and when the photodiode turns on it merely specifies an actual repertoire where states (00,01,10,11) have, respectively, probability (0,0,1/2,1/2). This corresponds in Q to a single q-arrow, one bit long, going from the potential, maximum entropy repertoire (1/4,1/4,1/4,1/4) to (0,0,1/2,1/2). Now imagine the light sensor is substituted by a temperature sensor with the same threshold and dynamic range—we have a thermistor rather than a photodiode. Although the physical device has changed, according to the IIT the experience, minimal as it is, has to be the same, since the informational relationship that is generated by the two devices is identical. Similarly, an AND gate when silent and an OR gate when firing also generate the same shape in Q, and therefore must generate the same minimal experience (it can be shown that the two shapes are isomorphic, that is, have the same symmetries; Balduzzi and Tononi, unpubl.). In other words, different "physical" systems (possibly in different states) generate the same experience if the shape of the informational relationships they specify is the same. On the other hand, more complex networks of causal interactions are likely to create highly idiosyncratic shapes, so systems of high Φ are unlikely to generate exactly identical experiences.

If experience is integrated information, it follows that only the informational relationships within a complex (those that give the quale its shape) contribute to experience. Conversely, the informational relationships that exist outside the main complex—for example, those involving sensory afferents or cortico-subcortical loops implementing informationally insulated subroutines—do not make it into the quale, and therefore do not contribute either to the quantity or to the quality of consciousness.

Note also that informational relationships, and thus the shape of the quale, are specified both by the elements that are firing and by those that are not. This is natural considering that an element that does not fire will typically rule out some previous states of affairs (those that would have made it fire), and thereby it will contribute to specifying the actual repertoire. Indeed, many silent elements can rule out, in combination, a vast number of previous states and thus be highly informative. From a neurophysiological point of view, such a corollary may lead to counterintuitive predictions. For example, take elements (neurons) within the main complex that happen to be silent when one is having a particular experience. If one were to temporarily disable these neurons (*e.g.,* make them *incapable* of firing), the prediction is that, though the system state (firing pattern) would remain the same, the quantity and quality of experience would change (Tononi, 2004; Balduzzi and Tononi, 2008).

It is important to see what Φ corresponds to in this representation (Fig. 7A). The minimum information partition (MIP) is just another point in Q: the one specified by the connections *within* the minimal parts only, leaving out the contribution of the connections *among* the parts. This point is the actual repertoire corresponding to the product of the actual repertoires of the parts taken independently. Φ corresponds then to an arrow linking this point to the top of the solid. In this view, the q-edges leading to the minimum information bipartition provide the natural "base" upon which the solid rests—the informational relationships generated *within* the parts upon which are built the informational relationships *among* the parts. The Φ-arrow can then be thought of as the height of the solid—or rather, to employ a metaphor, as the highest pole holding up a tent. For example, if Φ is zero (say a system decomposes into two independent complexes as in Fig. 7B), the tent corresponding to the system is flat—it has no shape—since the actual repertoire of the system collapses onto its base (MIP). This is precisely what it means when $\Phi = 0$. Conversely, the higher the Φ value of a complex (the higher the tent or solid), the more "breathing room" there is for the various informational relationships within the complex (the edges of the solid or the seams of the tent) to express themselves.

In summary, and not very rigorously, the generation of an experience can be thought of as the erection of a tent with a very complex structure: the edges are the tension lines generated by each subset of connections (the respective q-arrow or informational relationship). The tent literally takes shape when the connections are engaged and specify actual repertoires. Perhaps an even more daring metaphor would be the following: whenever the mechanisms of a complex unfold and specify informational relationships, the flower of experience blooms.

From phenomenology to geometry

The notions just sketched aim at providing a framework for translating the seemingly ineffable qualitative properties of phenomenology into the language of mathematics, specifically, the language of informational relationships (q-arrows) in Q. Ideally, when sufficiently developed, such language should permit the geometric characterization of phenomenological properties generated by the human brain. In principle, it should also allow us to characterize the phenomenology of other systems. After all, in this framework the experience of a bat echo-locating in a cave is just another shape in Q and, at least in principle, shapes can be compared objectively.

At present, due to the combinatorial problems posed by deriving the shape of the quale produced by systems of just a few elements, and to the additional difficulties posed by representing such high-dimensional objects, the best one can hope for is to show that the language of Q can capture, in principle, some of the basic distinctions that can be made in our own phenomenology, as well as some key neuropsy-

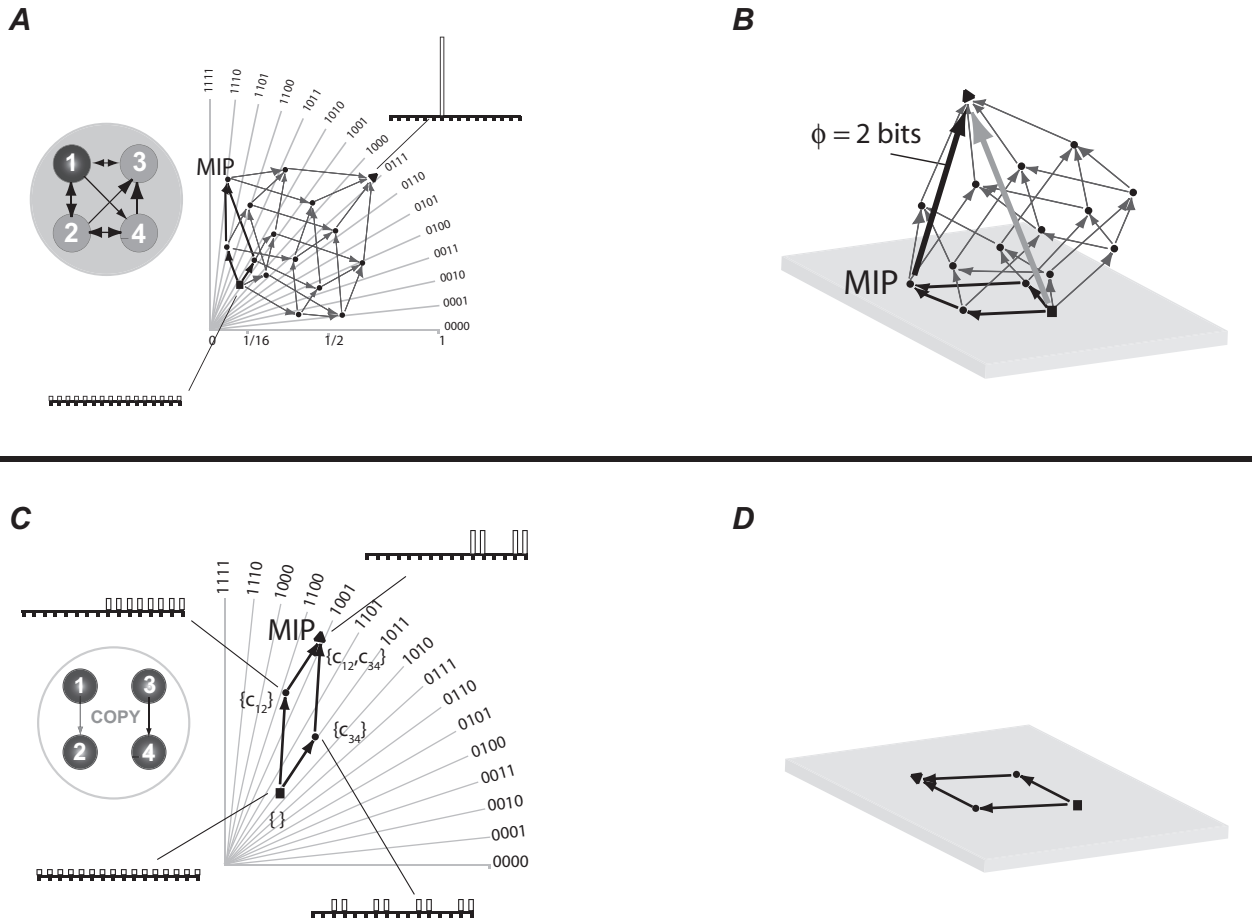

Figure 7. The tent analogy. (A) The system of Fig. 2A′ / Fig. 5. (B) The q-edges converging on the minimum information partition of the system (MIP) form the natural base on which the complex rests, depicted as a "tent." The informational relationships *among* the parts are built on top of the informational relationships generated independently *within* the minimal parts. From this perspective the Φ q-arrow (in black) is simply the tent pole holding the quale up above its base; the length (divergence) of the pole expresses the breathing room in the system. The thick gray q-arrow represents the information generated by the entire system. (C) The system of Fig. 2A. The quale (not) generated by the two photodiodes considered as a single system. As shown in Fig. 2A, the system reduces to two independent parts, so it does not exist as a single entity. (D) Note that in this case the quale reduces to the MIP: the "tent" collapses onto its base, so there is no breathing room for informational relationships within the system. The quale generated by each part considered in isolation does exist, corresponding to an identical q-arrow for each couple.

chological observations (Balduzzi and Tononi, unpubl.). A short list includes the following:

(i) Experience is divided into modalities, like the classic senses of sight, hearing, touch, smell, and taste (and several others), as well as submodalities, like visual color and visual shape. What do these broad distinctions correspond to in Q? According to the IIT, modalities are sets of densely entangled q-arrows (*modes*) that form distinct *sub-shapes* in the quale; submodalities are subsets of even more densely entangled q-arrows (sub-modes) within a larger mode, thus forming distinct sub-sub-shapes (Fig. 8). As a two-dimensional analog, imagine a given multimodal experience as the shape of the three-continent complex constituted by Europe, Asia, and Africa. The three continents are distinct sub-shapes, yet they are all part of the same landmass, just as modalities are parts of the same consciousness. Moreover, within each continent there are peninsulas (sub-sub-shapes), like Italy in Europe, just as there are submodalities within modalities.

(ii) Some experiences appear to be "elementary," in that they cannot be further decomposed. A typical example is what philosophers call a "quale" in the narrow sense—say a pure color like red, or a pain, or an itch: it is difficult, if not impossible, to identify any further phenomenological structure within the experience of red. According to the IIT, such elementary experiences correspond to sub-modes that do not contain any more densely entangled sub-sub-modes (*elementary modes*, Fig. 8).

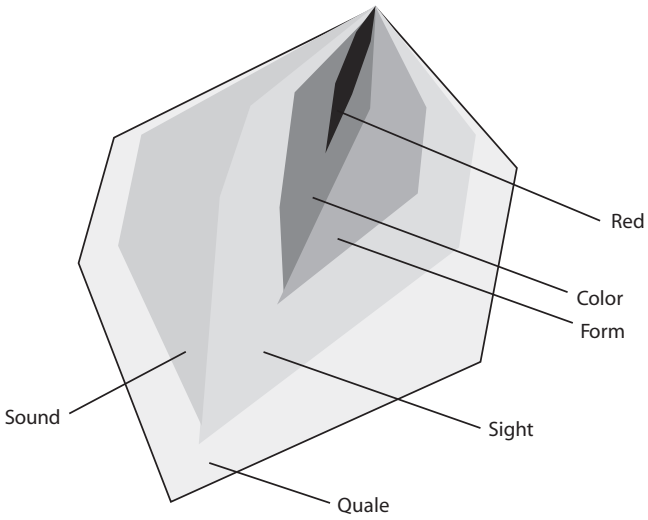

Figure 8. Modes. Schematic depiction of modes and sub-modes. A mode, indicated by a polygon within the quale (light gray with black border), is a set of q-arrows that are more densely entangled than surrounding q-arrows, and can be considered as clusters of informational relationships constituting distinctive "sub-shapes" in Q. Two different modes could correspond, for example, to the modalities of sight and sound. A sub-mode within a mode is a set of q-arrows that is even more densely entangled (a sub-sub-shape in Q). Color and form could correspond to two sub-modes within the visual mode. The thin black polygon represents an elementary mode, which does not contain more densely entangled q-arrows. Elementary modes could correspond to experiential qualities that cannot be further decomposed, such as the color "red" (qualia in the narrow sense.)

(iii) Some experiences are homogeneous and others are composite: for example, a full-field experience of blue, as when watching a cloudless sky, compared to that of a busy market street. In Q, homogeneous experiences translate to a single homogeneous shape, and composite ones into a composite shape with many distinguishable sub-shapes (modes and sub-modes).

(iv) Some experiences are hierarchically organized. Take seeing a face: we see at once that as a whole it is somebody's face, but we also see that it has parts such as hair, eyes, nose, and mouth, and that those are made in turn of specifically oriented segments. The subjective experience is constructed from informational relationships (q-arrows) that are entangled (not reducible to a product of independent components) across hierarchical levels. For example, informational relationships constituting "face" would be more densely tangled than unnatural combinations such as seen in certain Cubist paintings. The sub-shape of the quale corresponding to the experience of seeing a face is then an overlapping hierarchy of tangled q-arrows, embodying relationships within and across levels.

(v) We recognize intuitively that the way we perceive taste, smell, and maybe color, is organized phenomenologically in a "categorical" manner, quite different from, say, the "topographical" manner in which we perceive space in vision, audition, or touch. According to the IIT, these hard-to-articulate phenomenological differences correspond to different basic sub-shapes in Q, such as 2^n-dimensional *grid-like* structures and *pyramid-like* structures, which emerge naturally from the underlying neuroanatomy.

(vi) Some experiences are more alike than others. Blue is certainly different from red (and irreducible to red), but clearly it seems even more different from middle C on the oboe. In the IIT framework, in Q colors correspond to different sub-shapes of the same kind (say pyramids pointing in different directions) and sounds to very different sub-shapes (say tetrahedra). In principle, such subjective similarities and differences can be investigated by employing objective measures of *similarity between shapes* (*e.g.,* considering the number and kinds of symmetries involved in specifying shapes that are generated in Q by different neuroanatomical circuits).

(vii) Experiences can be refined through learning and changes in connectivity. Suppose one learns to distinguish wine from water, then red wines from whites, then different varietals. Presumably, underlying this phenomenological refinement is a neurobiological refinement: neurons that initially were connected indiscriminately to the same afferents become more specialized and split into sub-groups with partially segregated afferents. This process has a straightforward equivalent in Q: the single q-arrow generated initially by those afferents *splits* into two or more q-arrows pointing in different directions, and the overall sub-shape of the quale is correspondingly refined.

(viii) Qualia in the narrow sense (elementary modes) exist "at the top of experience" and not at its bottom. Consider the experience of seeing a pure color, such as red. The evidence suggests that the "neural correlate" (Crick and Koch, 2003) of color, including red, is probably a set of neurons and connections in the fusiform gyrus, maybe in area V8 (ideally, neurons in this area are activated whenever a subject sees red and not otherwise, if stimulated trigger the experience of red, and if lesioned abolish the capacity to see red). Certain achromatopsic subjects with dysfunctions in this general area seem to lack the feeling of what it is like to see color, its "coloredness," including the "redness" of red. They cannot experience, imagine, remember, or even dream of color, though they may talk about it, just as we could talk about echolocation, from a third-person perspective (van Zandvoort *et al.,* 2007). Contrast such subjects, who are otherwise perfectly conscious, with vegetative patients, who are for all intents and purposes unconscious. Some of these patients may show behavioral and neurophysiological evidence for residual function in an isolated brain area (Posner and Plum, 2007). Yet it seems highly unlikely that a vegetative patient with residual activity exclusively in V8 should enjoy the vivid perceptions of color just as we do, while being otherwise unconscious.

The IIT provides a straightforward account for this difference. To see how, consider again Figure 6A: call "r" the

connections targeting the "red" neurons in V8 that confer them their selectivity, and non-r (¬r) all the other connections within the main corticothalamic complex. Adding r in isolation at the bottom of Q (null context) yields a small q-arrow (called the *down-set of red* or ↓ r) that points in a direction representing how r by itself shapes the maximum entropy distribution into an actual repertoire. Schematically, this situation resembles that of a vegetative patient with V8 and its afferents intact but the rest of the corticothalamic system destroyed. The shape of the experience or quale reduces to this q-arrow, so its quantity is minimal (Φ for this q-arrow is obviously low) and its quality minimally specified: as we have seen with the photodiode, r by itself cannot specify whether the experience is a color rather than something else such as a shape, whether it is visual or not, sensory or not, and so on.

By contrast, subtract r from the set of all connections, so one is left with ¬r. This "lesion" collapses the q-fold specified by r in all contexts, including the q-arrow, called the *up-set of non-red* (↑ ¬r), which starts from the full context provided by all other connections ¬r and reaches the top of the quale.[10] This q-arrow will typically be much longer and point in a different direction than the q-arrow generated by r at the bottom of the quale. This is because, the fuller the context, the more r can shape the actual repertoire. Schematically, removing r from the top resembles the situation of an achromatopsic patient with a selective lesion of V8: the bulk of the experience or quale remains intact (Φ remains high), but a noticeable feature of its shape collapses (the upset of non-red). According to the IIT, the feature of the shape of the quale specified by "the upset of non-red" captures the very quality or "redness" of red.[11]

It is worth remarking that the last example also shows why specific qualities of consciousness, such as *the "redness" of red, while generated by a local mechanism, cannot be reduced to it*. If an achromatopsic subject without the r connections lacks precisely the "redness" of red, whereas a vegetative patient with just the r connections is essentially unconscious, then the redness of red cannot map directly to the mechanism implemented by the r connections. However, the redness of red can map nicely onto the informational relationships specified by r, as these change dramatically between the null context (vegetative patient) and the full context (achromatopsic subject).

A Provisional Manifesto

To recapitulate, the IIT claims that the quantity of consciousness is given by the integrated information (Φ) generated by a complex of interacting elements, and its quality by the shape in Q specified by their informational relationships. As I have tried to indicate here, this theoretical framework can account for basic neurobiological and neuropsychological observations. Moreover, the same framework can be extended to begin translating phenomenology into the language of mathematics.

At present, the very notion of a theoretical approach to consciousness may appear far-fetched, yet the nature of the problems posed by a science of consciousness requires a combination of experiment and theory: one could say that theories without experiments are lame, but experiments without theories are blind. For instance, only a theoretical framework can go beyond a provisional list of candidate mechanisms or brain areas and provide a principled explanation of why they may be relevant. Also, only a theory can account, in a coherent manner, for key but puzzling facts about consciousness and the brain, such as the association of consciousness with the corticothalamic but not the cerebellar system, the "unconscious" functioning of many corticosubcortical circuits, or the fading of consciousness during certain stages of sleep or epilepsy.

A theory should also generate relevant corollaries. For example, the IIT predicts that consciousness depends exclusively on the ability of a system to generate integrated information: whether or not the system is interacting with the environment on the sensory and motor side, it deploys language, capacity for reflection, attention, episodic memory, a sense of space, of the body, and of the self. These are obviously important functions of complex brains and help shape its connectivity. Nevertheless, contrary to some common intuitions, but consistent with the overall neurological evidence, none of these functions seems absolutely necessary for the generation of consciousness "here and now" (Tononi and Laureys, 2008).

Finally, a theory should be able to help in "difficult" cases that challenge our intuition or our standard ways to assess consciousness. For instance, the IIT says that the presence and extent of consciousness can be determined, in principle, also in cases in which we have no verbal report, such as infants or animals, or in neurological conditions such as minimally conscious states, akinetic mutism, psychomotor seizures, and sleepwalking. In practice, of course, measuring Φ accurately in such systems will not be easy, but approximations and informed estimates are certainly conceivable. Whether these and other predictions turn out to be compatible with future clinical and experimental evidence, a coherent theoretical framework should at least help to systematize a number of neuropsychological and neurobiological results that might otherwise seem disparate (Albus *et al.,* 2007).

In the remaining part of this article, I briefly consider some implications of the IIT for the place of experience in our view of the world.

Consciousness as a fundamental property

According to the IIT, consciousness is one and the same thing as integrated information. This identity, which is

predicated on the phenomenological thought experiments at the origin of the IIT, has ontological consequences. Consciousness exists beyond any doubt (indeed, it is the only thing whose existence is beyond doubt). If consciousness is integrated information, then integrated information exists. Moreover, according to the IIT, it exists as a fundamental quantity—as fundamental as mass, charge, or energy. As long as there is a functional mechanism in a certain state, it must exist *ipso facto* as integrated information; specifically, it exists as an experience of a certain quality (the shape of the quale it generates) and quantity (its "height" Φ).[12]

If one accepts these premises, a useful way of thinking about consciousness as a fundamental property is as follows. We are by now used to considering the universe as a vast empty space that contains enormous conglomerations of mass, charge, and energy—giant bright entities (where brightness reflects energy or mass) from planets to stars to galaxies. In this view (that is, in terms of mass, charge, or energy), each of us constitutes an extremely small, dim portion of what exists—indeed, hardly more than a speck of dust.

However, if consciousness (*i.e.,* integrated information) exists as a fundamental property, an equally valid view of the universe is this: a vast empty space that contains mostly nothing, and occasionally just specks of integrated information (Φ)—mere dust, indeed—even there where the mass-charge–energy perspective reveals huge conglomerates. On the other hand, one small corner of the known universe contains a remarkable concentration of extremely bright entities (where brightness reflects high Φ), orders of magnitude brighter than anything around them. Each bright "Φ-star" is the main complex of an individual human being (and most likely, of individual animals).[13] I argue that such Φ-centric view is at least as valid as that of a universe dominated by mass, charge, and energy. In fact, it may be more valid, since to be highly conscious (to have high Φ) implies that there is something it is like to be you, whereas if you just have high mass, charge, or energy, there may be little or nothing it is like to be you. From this standpoint, it would seem that entities with high Φ *exist* in a stronger sense than entities of high mass.

Intriguingly, it has been suggested, from a different perspective, that information may be, in an ontological sense, prior to conventional physical properties (the *it from bit* perspective; Wheeler and Ford, 1998). This may well be true but, according to the IIT, only if one substitutes "integrated information" for information.[14] Information that is not integrated, I have argued, is not associated with experience, and thus does not really exist as such: it can only be given a vicarious existence by a conscious observer who exploits it to achieve certain discriminations within his main complex. Indeed, the same "information" may produce very different consequences in different observers, so it only exists through them but not in and of itself.

Consciousness as an intrinsic property

Consciousness, as a fundamental property, is also an *intrinsic* property. This simply means that a complex generating integrated information is conscious in a certain way regardless of any extrinsic perspective. This point is especially relevant if we consider how difficult it is to measure the quantity of integrated information, not to mention the shape of a quale, for any realistic system. If we want to know what are the borders of a certain complex, the amount of integrated information it generates, the set of informational relationships it specifies, and the spatio-temporal grain at which Φ is highest (see below), we need to perform a prohibitively large set of computations. One would need to perturb a system in all possible ways and use Bayes' rule to keep track of the probabilities of the previous states given the current output, and then calculate the relative entropy between the potential and the actual distributions. Moreover, this must be done for all possible subsets of a system (to find complexes) and for all combinations of connections (to obtain the shape of each quale). Finally, the calculations must be repeated at multiple spatial and temporal scales to determine what is the optimal grain size, in space and time, for generating integrated information (see below). It goes without saying that these calculations are presently unfeasible for anything but the smallest systems. It also goes without saying that a complex itself cannot and need not go through such calculations: it is intrinsically conscious in this or that way. In fact, it needs as little to "calculate" all the relevant probability distributions to generate consciousness and specify its quality, as a body of a certain mass needs to "calculate" how much gravitational mass it has in order to attract other bodies.

Another way to express this aspect of integrated information is to say that consciousness can be characterized extrinsically as a disposition or *potentiality* –in this case as the potential discriminations that a complex can do on its possible states, through all combinations of its mechanisms, yet from an intrinsic perspective it is undeniably *actual*. While this may sound strange, fundamental quantities associated with physical systems can also be characterized as dispositions or potentialities, yet have actual effects. For example, mass can be characterized as a potentiality—say the resistance that a body would offer to acceleration by a force—yet it exerts undeniably actual effects, such as actually attracting other masses if these turn out to be there. Similarly, a mechanism's potential for integrated information becomes actual by virtue of the fact that the mechanism is actually in a particular state. Paraphrasing E. M. Forster, one could express this fact as follows: *How do I know what I am till I see what I do?*

Being and describing

According to the IIT, a full *description* of the set of informational relationships generated by a complex at a given time should say all there is to say about the experience it is having at that time: nothing else needs to be added.[17] Nevertheless, the IIT also implies that to be conscious—say to have a vivid experience of pure red—one needs to *be* a complex of high Φ; there is no other way. Obviously, although a full description can provide understanding of what experience is and how it can be generated, it cannot substitute for it: *being is not describing*. This point should be uncontroversial, but it is worth mentioning because of a well-known argument against a scientific explanation of consciousness, best exemplified by a thought experiment involving Mary, a neuroscientist in the 23rd century (Jackson, 1986). Mary knows everything about the brain processes responsible for color vision, but has lived her whole life in a black-and-white room and has never seen any color.[18] The argument goes that, despite her complete knowledge of color vision, Mary does not know what it is like to experience a color: it follows that there is some knowledge about conscious experience that cannot be deduced from knowledge about brain processes. The argument loses its strength the moment one realizes that consciousness is a way of being rather than a way of knowing. According to the IIT, being implies "knowing" from the inside, in the sense of generating information about one's previous state. Describing, instead, implies "knowing" from the outside. This conclusion is in no way surprising: just consider that though we understand quite well how energy is generated by atomic fission, unless atomic fission occurs, no energy is generated—no amount of description will substitute.

Observer pitfalls: minimal elements and minimal interactions

Because integrated information is an intrinsic property, it is especially important that one avoid the observer fallacy in estimating how much of it is generated by a system. Consider the system in Figure 9A (top). An observer might assume that the system is made up of two units, each with a repertoire of 2^n states. If the lower unit copies the output of the upper unit, then this two-unit system generates n bits of integrated information—it would seem trivial to implement systems with arbitrarily large values of Φ. But how is the system really built? Figure 9A (bottom) shows a possible architecture: each "unit" is actually not a unit at all, but it contains n binary elements. Each upper element is then connected to the corresponding lower element. Seen this way, it becomes obvious that the system is not a complex generating n bits of integrated information, but rather a collection of independent couples (or photodiodes) each generating 1 bit of integrated information, just as in Figure 2. Note that, if we try to "integrate" the couples by adding horizontal connections between elements, we reduce the available information. Thus, integrated information has to be evaluated from the perspective of the system itself, starting from its elementary, indivisible components (see also the next point), and not by arbitrarily imposing "units" from the perspective of an observer.

Figure 9B (top) illustrates a similar problem with respect to elementary operations. The system contains $n+1$ binary components, with a single component receiving inputs from the other n; the component fires if all n inputs are active. The minimum information partition is the total partition $P = \{X\}$ and $\Phi = n$ bits when the top component is firing, since it uniquely specifies the prior state of the other n components. Increasing the number of inputs feeding into the top component while maintaining the same rule—fire if and only if all inputs are active—seems to provide a method for constructing systems with high Φ[15] using binary components and a basic architecture that is certainly easy to describe. The difficulty once again lies in physically implementing a component that processes n inputs at a single point in space and at a single instant in time for large n. Figure 9B (bottom) shows a possible internal architecture of the component, constructed using a hierarchy of logical AND-gates. When analyzed at this level, it is apparent that the system generates 1 bit of integrated information regardless of the number of inputs that feed into the top component, since the bipartition framed by the dashed cut forms a bottleneck. As in the previous example, integrated information has to be evaluated from the perspective of the system itself, based on the elementary causal interactions its elements can perform, and not by arbitrarily imposing "rules" from the perspective of an observer with no regard to their actual implementation. It is well known that all computations (or Boolean functions) can be performed by elementary logical gates such as NOR or NAND gates acting on elementary binary elements. In principle, then, a system should be decomposed into minimal elements and minimal interactions—as elementary as they come in terms of physical implementation—before any pronouncement is made on its capacity to generate integrated information and thereby consciousness.[16]

Consciousness and the spatiotemporal grain of reality

An outstanding issue is finding a principled way to determine the proper spatial and temporal scale to measure informational relationships and integrated information. What are the elements upon which probability distributions of states are to be evaluated? For example, are they minicolumns or neurons? And what about molecules, atoms, or subatomic particles? Similarly, what is the "clock" to use to identify system states? Does it run in seconds, hundreds of milliseconds, milliseconds, or microseconds?

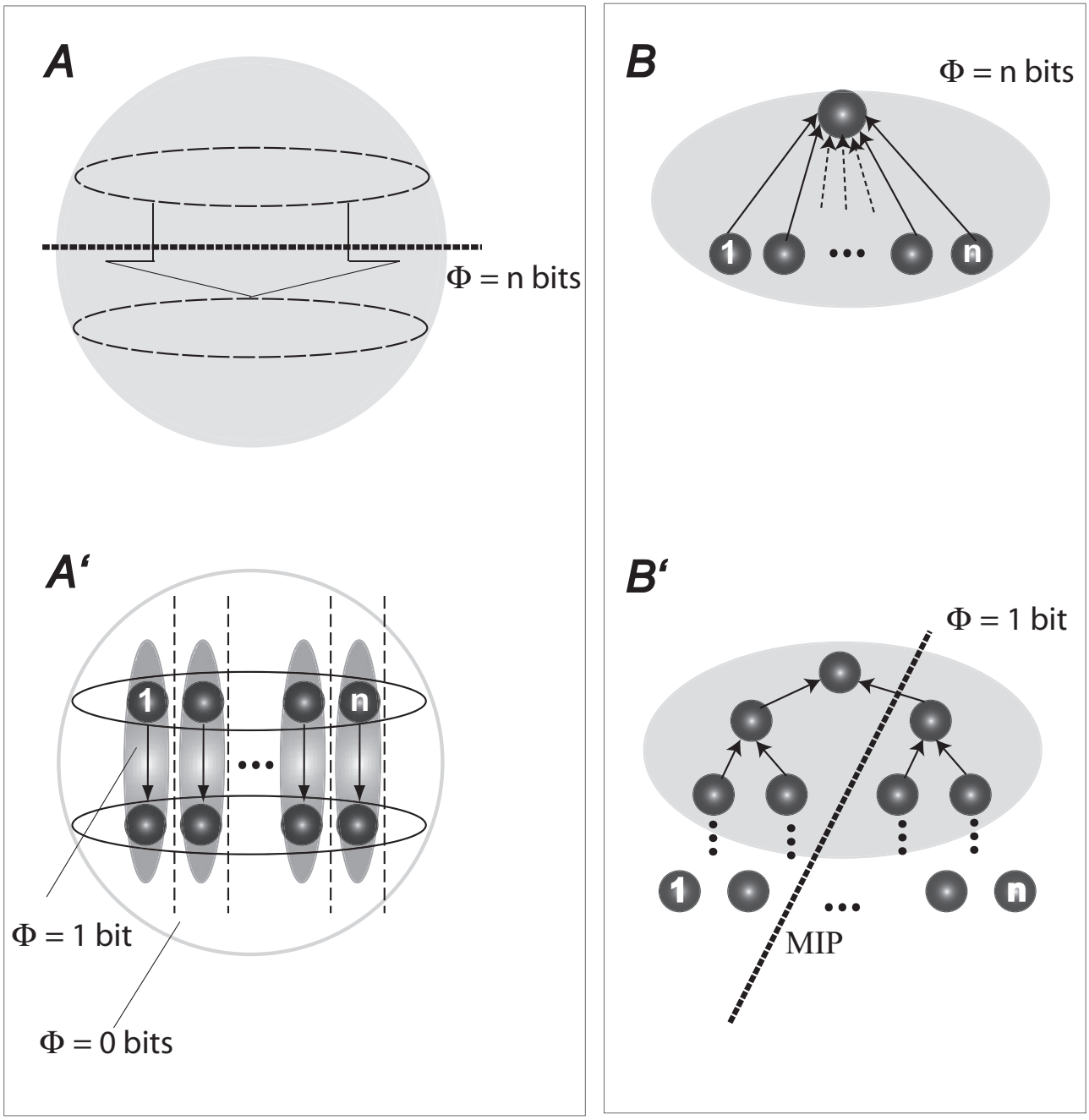

Figure 9. Analyzing systems in terms of elementary components and operations. (A) and (B) show systems that on the surface appear to generate a large amount of integrated information. The units in (A) have a repertoire of 2^n outputs, with the bottom unit copying the top. Integrated information is n bits. By analyzing the internal structure of the system in (A′) we find n disjoint couples, each integrating 1 bit of information; the entire system, however, is not integrated. (B) shows a system of binary units. The top unit receives inputs from eight other units and performs an *AND*-gate like operation, firing if and only if all eight inputs are spikes. Increasing the number of inputs appears to easily increase Φ without limit. (B′) examines a possible implementation of the internal architecture of the top unit using binary *AND*-gates. The architecture has a bottleneck, shown as the MIP line, so that $\Phi = 1$ bit regardless of the number of input units.

Properly addressing this issue requires a comprehensive theoretical approach to the relationship between integrated information, emergence, and memory (Balduzzi and Tononi, unpubl.). The working hypothesis is as follows (Tononi, 2004): In general, for any system, integrated information is generated at multiple spatiotemporal scales. In

particular, however, there will often be a privileged spatiotemporal "grain size" at which a given system forms a complex of highest Φ—the spatiotemporal scale at which it "exists" the most in terms of integrated information, and therefore of consciousness.

For example, while in the brain there are many more atoms than neurons, it is likely that complexes at the spatial scale of atoms are exceedingly small, or at any rate that they cannot maintain both functional specialization and long-range integration, thus yielding low values of Φ. At the other extreme, the spatial scale of cortical areas is almost certainly too coarse for yielding high values of Φ. Somewhere in between, most naturally at the grain size of neurons or minicolumns, the neuroanatomical arrangement ensures an ideal mix of functional specialization and integration, leading to the formation of a large complex of high Φ.

Similarly, with respect to time, neurons would yield zero Φ at the scale of microseconds, since there is simply not enough time for engaging their mechanisms. At long time scales, say hours, Φ would also be low, as output states would bear little relationship to input states. Somewhere in between, at a time scale of tens to hundreds of milliseconds, the firing pattern of a large complex of neurons should be maximally predictive of its previous state, thus yielding high Φ. It is not by chance, according to the IIT, that this is both the time scale at which experience seems to flow (Bachmann, 2000) and that at which long-range neuronal interactions occur (Dehaene *et al.,* 2003; Koch, 2004).[21]

This working hypothesis also suggests that the generation of integrated information may set an intrinsic framework for both space and time. With respect to time, for example, consider a complex generating a certain shape in Q through a fast mechanism, and another complex that generates exactly the same shape, but through a slower mechanism. It would seem that these two complexes should generate exactly the same experience, except that time would flow faster in one case and slower in the other. Similar considerations may apply to space. Also, according to the IIT, what constitutes a "state" of the system is not an arbitrary choice from an extrinsic perspective, but rather the spatiotemporal grain size at which the system can best generate information about its past: *what is, is what can make a difference*.

Consciousness as a graded quantity

The IIT claims that consciousness is not an all-or-none property, but is graded: specifically, it increases in proportion to a system's repertoire of discriminable states. Strictly speaking, then, the IIT implies that even a binary photodiode is not completely unconscious, but rather enjoys exactly 1 bit of consciousness. Moreover, the photodiode's consciousness has a certain quality to it—the simplest possible quality—that is captured by a single q-arrow of length 1 bit.[19]

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 (in fact, it does not exist as an entity). Moreover, panpsychism hardly has a solid conceptual foundation. The attribution of consciousness to all kinds of things is based more on an attempt to avoid dualism than on a principled analysis of what consciousness is. Similarly, panpsychism offers hardly any guidance as to what would determine the amount of consciousness associated with different things (such as humans, animals, plants, or rocks), or with the same thing at different times (say wakefulness and sleep), not to mention that it says nothing about what would determine the quality of experience.

A more relevant issue is the following: How can the theory attribute consciousness (albeit minimal) to a photodiode, while acknowledging that we "lose" consciousness every night when falling into dreamless sleep? After all, the sleeping brain likely generates more integrated information than a photodiode. Two considerations are in order. First, we have first-hand "experience" that consciousness can be graded: falling asleep is often a rapid process but, before we are "gone" altogether, we occasionally do go through some degree of restriction in the field of consciousness, where we are progressively less aware of ourselves and the environment. Something similar also happens at certain stages of alcohol intoxication. So the level of consciousness can indeed change around our typical waking baseline, allowing for some gradation.

Below a certain level of consciousness, however, it truly feels as if we fade away completely. But is consciousness really annihilated? Is it likely that when we "lose" consciousness the amount of integrated information generated by the corticothalamic main complex decreases nonlinearly? Computer simulations indicate that when the overall activation of corticothalamic networks goes below a certain level, there is a sudden drop in the average effective information between distant parts of the cortex (Tononi, unpubl. obs.). In other words, below a certain threshold of activation the corticothalamic system breaks down into nearly independent pieces and cannot sustain integrated patterns of firing. This could explain why it feels as if consciousness is vanishing in an almost all-or-none manner rather than diminishing progressively.[20]

The limited capacity of consciousness

It is often stated that the brain discards most of the incoming information, and that only a very small portion trickles into consciousness. Thus, though the retina can transmit millions of bits per second, some estimates suggest that just a few bits per second make it to consciousness (Nørretranders, 1998), which is abysmally little by engineering standards. Indeed, as shown by classic experiments, we cannot keep in mind more than a few things at a time.

For the IIT, however, the informativeness of consciousness is not related to how many chunks of information a single experience might contain. Instead, it relates to how many different states are ruled out. Since we can easily discriminate among trillions of conscious states within a fraction of a second, the informativeness of conscious experience must be considerable. Presumably, the so-called capacity limitation of consciousness reflects an upper bound on how many partially independent subprocesses can be sustained within the main complex without compromising its integration.

Another consequence of the need for integration is the seemingly serial nature of consciousness. Since a complex constitutes a single entity, it must move from one global state to another, and its temporal evolution must follow a single trajectory. Indeed, dual-task paradigms and the psychological refractory period show that decisions or choices can only occur one at a time (Pashler, 1998). Such choices take around 150 milliseconds, a figure remarkably close to the lower limit of the time typically needed for conscious integration.

More generally, although transmitting and storing information is relatively cheap and easy, generating integrated information would seem to be more expensive and difficult. Ensuring that a system forms a complex (integration) requires many connections per element, and connections are usually expensive. At the same time, ensuring that the complex can discriminate among a large number of states (information) requires that connections are patterned so that elements are both functionally specialized and capable of acting as a single entity, which is usually difficult. Thus, it may be more fitting to say that the brain, rather than discarding information, sifts through the chaff to extract precious kernels of integrated information. To use another metaphor, if information were like carbon, mere information would be like a heap of coal, and integrated information like a precious diamond.

Conscious artifacts?

Many scientists think that other species beyond humans are likely to be conscious (Koch, 2004) based on commonalities of behavior and on the overall similarity between their corticothalamic system and ours. But when it comes to species that have radically different neural organization, such as fruit flies, or even more when one considers man-made artifacts, arguments from analogy lose their strength, and it is hard to know what to think. The IIT has a straightforward position on this issue: to the extent that a mechanism is capable of generating integrated information, no matter whether it is organic or not, whether it is built of neurons or of silicon chips, and independent of its ability to report, it will have consciousness. Thus, the theory implies that it should be possible to construct highly conscious artifacts by endowing them with a complex of high Φ (Koch and Tononi, 2008). Moreover, it should be possible to design the quality of their conscious experience by appropriately structuring their effective information matrix.

Such a position should not be read as implying that building conscious artifacts may be easy, or that many existing man-made products, especially "complicated" ones, should be expected to have high values of Φ. The conditions needed to build complexes of high Φ, such as a combination of functional specialization and integration, are apparently not easy to achieve. Moreover, computer simulations suggest that seemingly "complicated" networks with many nodes and connections, whose connection diagram superficially suggests a high level of "integration," usually turn out to break down into small local complexes of low Φ, or to form a single entity with a small repertoire of states and therefore also of low Φ: a paradigmatic example is a network with full connectivity, which can be shown to generate at most 1 bit of integrated information (Balduzzi and Tononi, 2008). Though we do not know how to calculate the amount of integrated information, not to mention the shape of the qualia, generated by structures such as a computer chip, the World Wide Web, or the proverbial network of Chinese talking on the phone (Block, 1978), it is likely that the same principles apply: high Φ requires a very special kind of complexity, not just having many elements intricately linked. Just think of something as complex as the cerebellum and its negligible contribution to consciousness.

Whether certain kinds of random networks (Tononi and Sporns, 2003), or even periodic network such as grids (Balduzzi and Tononi, 2008), could achieve high values of Φ (albeit inefficiently) by simply increasing the number of elements remains to be determined. The brain certainly exploits grid-like arrangements (as in early sensory areas) and certain kinds of near-random connectivity (as in prefrontal areas and perhaps, at a finer scale, everywhere else). Moreover, the small world architecture of the cerebral cortex and its hub-like backbone may be especially well-suited to integrating information (Sporns *et al.,* 2000; Hagmann *et al.,* 2008). At present, even for very small networks of just a dozen elements, the only way to increase Φ is by brute-force optimization, which is clearly unfeasible for more realistic networks, or through adaptation to a rich environment (Tononi *et al.,* 1996).

Consciousness and meaning

The notion of integrated information and, more generally, the set of informational relationships that constitute a quale, are closely related to the notion of meaning and, more generally, semantics. Here I briefly discuss how meaning requires a system capable of integrating information and, more specifically, how meaning is captured by concepts.

For the IIT, mechanisms generate meanings. Moreover, only the mechanisms within a single complex do so. A mechanism modifies a probability distribution (the context to which it is applied) into another distribution, thereby specifying an informational relationship. In essence, then, a mechanism rules out certain states and rules in others. Note the parallel with semantics, where a sentence's meaning is specified by the possible worlds in which it is true and false. Also, as in semantics, the meaning changes depending on the context in which the mechanism acts. For the IIT, however, meaning is only meaningful within a complex—mechanisms belonging to disjoint complexes do not generate meaning. In fact, what is meaningful is each individual experience, and its meaning is completely and univocally specified by the shape of its quale. For example, a photodiode[22] generating a single q-arrow means (*i.e.,* specifies) very little, whereas a large and complex quale means (*i.e.,* specifies) much more. The IIT is also precise about the possible worlds that need be considered: they are the states encompassed by the maximum entropy distribution of a complex. How meanings "in the head" of different subjects refer to the external world is a different matter, which requires considering the matching between internal and external relationships (see below).

Recall that *concepts* are entangled q-arrows that group together certain states of affairs in a way that cannot be decomposed into the mere sum of simpler groupings (see also Feldman, 2003). Figure 10 shows two systems comprising four input elements (sensors) and four output elements (detectors). The "copy" system (Fig. 10A, similar to the camera example in Fig. 2, left side) is such that each output element is connected to a different input element, implementing for each sensor-detector couple the function "D = S." The copy system relays all 4 bits in the input but, since it decomposes into four separate complexes, it generates no integrated information. Each sensor-detector couple generates 1 bit of integrated information and a single informational relationship (q-arrow), corresponding to the simplest possible concept: that things are one way rather than another way (just like the photodiode in Fig. 1).

Consider now the "conceptual" system (Fig. 10B). In this case, each output element receives connections from all four input elements, and performs a more complex Boolean function on the input.[23] For example, output element 5 could be implementing a "parity" function on the four input elements (it is on if an odd number of inputs are on, and off otherwise); element 6 a "symmetry" function (on if the arrangement of on-and-off inputs is symmetric); element 7 a "contiguity" function (on if on-or-off input elements are not separated by an element of the other sign); and element 8 a "balance" function (on if there are an equal number of on and off input elements).[24] In this case, the q-arrow generated by each output element (*i.e.,* by its afferent connec-

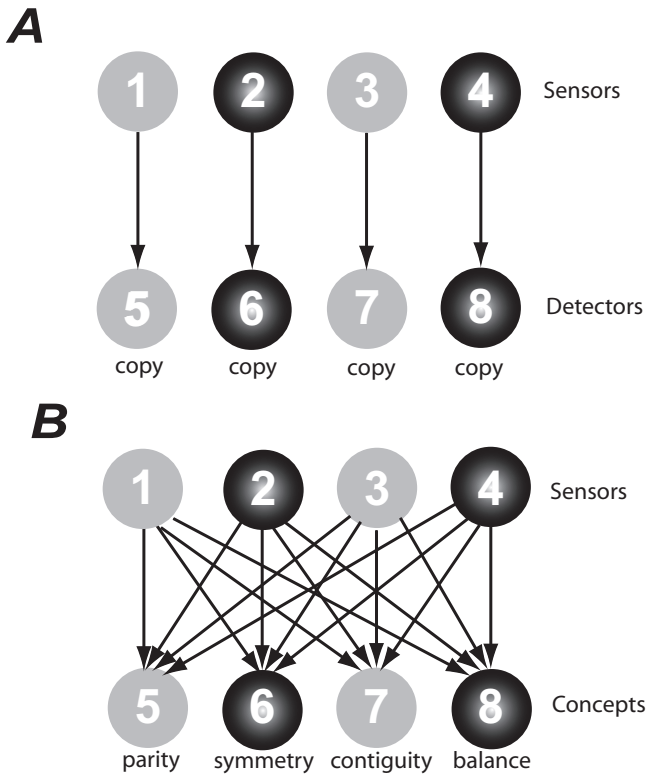

Figure 10. Meaning. (A) The "copy system." Each output element is connected to a different input element, implementing for each sensor-detector couple the function "D = S." The copy system relays all four bits in the input but, since it decomposes into four separate complexes, it generates no integrated information. Each sensor-detector couple generates 1 bit of integrated information and a single informational relationship (q-arrow), corresponding to the simplest possible concept: that things are one way rather than another way (just like the photodiode in Fig. 1). (B) The "conceptual" system. Each output element receives connections from all four input elements, and performs a more complex Boolean function on the input. The q-arrow generated by each output element (*i.e.,* by its afferent connections) is entangled (the information generated jointly by its four afferent connections is higher than the sum of the information generated by each connection independently). An entangled q-arrow constitutes a *concept*. In this case, the first element being off means "even" input, the second on means "symmetrical," the third off "non-contiguous," the fourth on "balanced." The q-arrow generated by all afferents to output elements considered together is also entangled, and means something like this: things are this particular way—an even, symmetrical, non-contiguous, balanced input—rather than many different ways. The conceptual system has literally added *meaning* to the input string. Moreover, the conceptual system realizes this concept as a single entity—a complex having high integrated information—rather than as a collection of smaller entities, each of which realizes only a partial concept.

tions) is entangled: the information generated jointly by its four afferent connections is higher than the sum of the information generated by each connection independently (for example, the parity function can only be computed when all inputs are considered together). As I mentioned above, an entangled q-arrow constitutes a *concept* in Q, here embodied in single output elements integrating globally over all four input elements. Moreover, in this case the four output elements specify different concepts, and thus generate information about different aspects of the input string.[25] Thus, the first element being off means "even" input, the second on means "symmetrical," the third off "non-contiguous," the fourth on "balanced." The q-arrow generated by all afferents to the output elements taken together is also entangled: the information generated jointly by all afferent connections is higher than the sum of the information generated independently by the afferents to each output element,[26] meaning something like this: things are this particular way—an even, symmetrical, non-contiguous, balanced input—rather than many different ways. The conceptual system has literally added *meaning* to the input string. Moreover, the conceptual system realizes this concept as a single entity—a complex having high integrated information—rather than as a collection of smaller entities, each of which realizes only a partial concept.

Indeed, meaning is truly in the eye of the beholder: an input string as such is meaningless, but becomes meaningful the moment it is "read" by a complex with a rich conceptual structure (corresponding to high Φ). Moreover, a complex with many different concepts will "read" meaning into anything, whether the meaning is there or not. It goes without saying that it is a good idea to build such complexes in such a way that its concepts are meaningful for interpreting the environment (for example, because they help predict future inputs). Finally, the more a system is able to conceptualize, the more it "understands"; or, if it was built to predict an environment, the more it "knows." Imagine that you do not know Chinese and are presented with a large number of Chinese characters. By and large, you will group them into the category (concept) of "must be something in Chinese," since they are all equivalent to you. After you have learned Chinese, however, each of the characters acquires a new, individual meaning (this one is a this, and that one is a that)—the input is the same, but the meaning has grown.[27]

The richness of qualia space

People often marvel at the immensity of the known universe, and wonder about other possible universes that we may never know. But perhaps even more awe-inspiring is the variety and complexity of nature around us. Just think of the number of different shapes that surround us, and their remarkable internal organization (see cover). This is certainly true of nonliving things, at multiple scales: think of crystals or, at a much grander scale, of mountains. But it is spectacularly true of living organisms, also at multiple scales: from the vast catalog of proteins and protein complexes—all of different shapes—to the inventory of cells, to that of organs, to the ramified tree of species, and within each species, to the panoply of different individuals. One could go on, and note how much of our own creations in engineering, science, and art also represent the generation of novel shapes, never seen before, again in astonishing variety. Perhaps most relevant in this context is to consider how even more extraordinary shapes would appear if we could look at them in more than just three dimensions and at the most appropriate level of organization. Take the brain at the synaptic level, and disentangle its connectional organization in all its complexity: if one could visualize the intricacy of the "connectome" (Sporns *et al.,* 2005) in a space of appropriate dimensionality, it would make for a remarkable shape indeed.

I mention all of this to come to a key aspect of the IIT: that experiences (*i.e.,* qualia) are shapes too. As remarkable as the "enchanted loom" of anatomical connectivity and firing patterns is, it pales compared to the shape of an experience in qualia space. For example, the complex generating the quale in Figure 5 has four elements (one of them firing) and nine connections among them. This simple system specifies a quale or shape that is described by 399 points in a 16-dimensional qualia space. It is hard to imagine what may be the complexity of the quale generated by a sizable portion of our brain. Add to this that the main complex within our brain, whatever its precise makeup in terms of neurons and connections, is presumably generating a different shape, just as remarkable, every few hundred milliseconds, often morphing smoothly into another shape as new informational relationships are specified through its mechanisms entering new states. Of course, we cannot dream of visualizing such shapes as qualia diagrams (we have a hard time with shapes generated by three elements). And yet, from a different perspective, we see and hear such shapes all the time, from the inside, as it were, since such shapes are actually the stuff our dreams are made of—indeed the stuff all experience is made of.

Consciousness and the world: matching informational relationships

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

passive, watching its own states change without having to act.[28] Depending on the informational relationships generated by its architecture, its qualia could be just as interesting as ours, whether or not they have anything to do with the causal architecture of the external world. Strange as this may sound, the theory says that it may be possible one day to construct a highly conscious, solipsistic entity.

Nevertheless, it is unlikely that a system having high Φ and interesting qualia would come to be by chance, but only by design or selection. Brain mechanisms, including those inside the main complex, are what they are by virtue of a long evolutionary history, individual development, and learning. Evolutionary history leads to the establishment of certain species-specific traits encoded in the genome, including brains and means to interact with the environment. Development and epigenetic processes lead to an appropriate scaffold of anatomical connections. Experience then refines neural connectivity in an ongoing manner though plastic processes, leading to the idiosyncrasies of the individual "connectome" and the memories it embeds.

Since for the IIT, experiences are informational relationships generated by mechanisms, what is the relationship between the structure of experience and the structure of the world? Again, this issue requires a comprehensive theoretical approach (Tononi *et al.,* 1996; Balduzzi and Tononi, unpubl.), but the main idea is simple enough. Through natural selection, epigenesis, and learning, informational relationships in the world mold informational relationships within the main complex that "resonate" best on a commensurate spatial and temporal scale. Moreover, over time these relationships will be shaped by an organism's values, to reflect relevance for survival. This process can be envisioned as the experiential analog of natural selection. As is well known, selective processes act on organisms through differential survival to modify gene frequencies (genotype), which in turn leads to the evolution of certain body forms and behaviors (extrinsic phenotype). Similarly, selective processes (Edelman, 1987) acting on synaptic connections through plastic changes modify brain mechanisms (neurotype), which in turn modifies informational relationships inside the main complex (intrinsic phenotype[29]) and thereby consciousness itself. In this way, qualia—the shapes of experience—come to be molded, sculpted, and refined by the informational structure of events in the world.

A working hypothesis is that the quantity of "matching" between the informational relationships inside a complex and the informational structure of the world can be evaluated, at least in principle, by comparing the value of Φ when a complex is exposed to the environment, to the value of Φ when the complex is isolated or "dreaming" (Tononi *et al.,* 1996). Similarly, the quality of matching can be evaluated by how the shapes of qualia "resonate" with the environment: for example, certain sub-shapes within a quale should "inflate" along certain dimensions when the complex is presented with appropriate stimuli.

This working hypothesis also suggests that morphogenesis and natural selection may be responsible for a progressive increase in the amount of integrated information generated by biological brains, and thus for the evolution of consciousness. This is because, in organisms exposed to a rich environment, plastic processes tend to increase functional specialization, while the brain's massive interconnectivity ensures neural and behavioral integration. In fact, it appears that as a system incorporates statistical regularities from its environment and learns to predict it, its capacity for integrated information may grow (Tononi *et al.,* 1996). It remains to be seen whether, based on the same principles, the construction of shapes even more extensive and complex may be achieved through nonbiological means.

Finally, the integrated information approach offers a straightforward perspective on why consciousness would be useful (Dennett, 1991). By definition, a highly conscious experience is a discrimination among trillions of alternatives—it specifies that what is the case is this particular state of affairs, which differs from a trillion other states of affairs in its own peculiar way, and in a way that is imbued with evolutionary value. Equivalently, one can say that a quale of high Φ represents a discrimination that is extremely context-sensitive, and thus likely to be useful. Experience is choice, and a highly conscious choice is a choice that is both highly informed and highly integrated.

Recall the photodiode. For it, turning on specifies that things are one way rather than another. What things might be like, it has 1 bit of a notion. For each of us, when the screen light turns on, the movie is about to begin.

Acknowledgments

I thank David Balduzzi, Chiara Cirelli, and Lice Ghilardi for their help, and the McDonnell Foundation for support.

Notes

[1] One could say that the theory starts from two basic phenomenological postulates—(i) experience is informative; (ii) experience is integrated—which are assumed to be immediately evident (or at least should be after going through the two thought experiments). In principle, the theory, including the mathematical formulation and its corollaries, should be derivable from these postulates.

[2] Note that two different distributions over the same states have relative entropy >0 even if they have the same entropy.

[3] One could paraphrase a classic definition of information (Bateson, 1972) and say that *information is a difference that made a difference* (the actual repertoire that can be discriminated by a given mechanism in a given state).

[4] In other words, *integrated information is a difference that made a difference to a system, to the extent that the system constitutes a single entity.*

[5] A phenomenon in which an observer may fail to perceive an image that is presented after a rapid succession of other images.

[6] A condition in which, when different images are presented to each eye, instead of seeing them superimposed, one perceives one image at a time, and which image one perceives switches every 2 seconds.

[7] The set of all subsets of connections forms a lattice (or more precisely a *logic,* characterized by an ordering relationship, join and meet operators, and a complement operator).

[8] Univocally implies, for example, that the "inverted spectrum" is impossible: a given shape (quale) specifies red and only red, another one green and only green. In turn, this implies that the neural mechanisms underlying the perception of red and green cannot be completely symmetric (Palmer, 1999).

[9] The set of all possible shapes generated by all possible systems corresponds to the set of all possible experiences.

[10] More precisely, the lesion collapses all q-arrows generated by r starting from any *context;* that is, it folds the quale along the *q-fold* specified by r.

[11] In lattices there is often a duality between elements (extensions) and attributes (intensions). Going up the lattice we move from elementary connections taken in isolation to all connections taken together. Going down the lattice, or up its dual, we move from the elementary attributes of a fully specified experience (the redness of red) to an undifferentiated experience, all of whose attributes are unspecified.

[12] In essence, the very existence of a functional mechanism in a given state is saying something like this: Given that I am a certain mechanism in good order, and that I am a certain state, things must have been this way, rather than other ways. In this sense, the information the mechanism generates is a statement about the universe made from its own intrinsic perspective—indeed, the only statement it can possibly make. Another way of saying this is that the mechanism is generating information by making an observation or *measurement*—where the mechanism is both the observer and the observed. In short, every (integrated) mechanism is an observer (of itself), and the state it is in is the result of that observation.

[13] There may be concentrations of such bright objects elsewhere in the universe, but at present we have no positive evidence.

[14] The notion of integrated information can in principle be extended to encompass quantum information. There are intriguing parallels between integrated information and quantum notions. Consider for example: (i) quantum superposition and the potential repertoire of a mechanism (in a sense, before it is engaged, a mechanisms exists in a superposition of all its possible output states); (ii) decoherence and the actual repertoire of a mechanism (when the mechanism is engaged and enters a certain state, it collapses the potential repertoire into the actual repertoire); (iii) quantum entanglement and integrated information (to the extent that one cannot perturb two elements independently, they are informationally one).

There are also some points of contact between the notion of integrated information and the approach advocated by relational quantum mechanics (Rovelli, 1996). The relational approach claims that system states exist only in relation to an observer, where an observer is another system (or a part of the same system). By contrast, the IIT says that a system can observe itself, though it can only do so by "measuring" its previous state. More generally, for the IIT, only complexes, and not arbitrary collections of elements, are real observers, whereas physics is usually indifferent to whether information is integrated or not.

Other interesting issues concern the relation between the conservation of information and the apparent increase in integrated information, and the finiteness of information (even in terms of qubits, the amount of information available to a physical system is finite). More generally, it seems useful to consider some of the paradoxes of information in physics from the intrinsic perspective, that is, as integrated information, where the observer is one and the same as the observed.

[15] Φ would be high for one specific firing pattern; for all other ones it would be very low.

[16] Here I ignore the issue of whether serial and parallel mechanisms are equivalent from the perspective of integrated information, as well as the issue of analog and digital computation (or quantum computation). In general, it must be asked to what extent two systems that are implemented differently actually specify the same complex and qualia when analyzed at the proper spatio-temporal grain.

[17] It is worth reiterating that a full description is practically out of the question for any realistic system.

[18] More appropriately, Mary should be like the achromatopsic patient mentioned above, since otherwise she might be able to dream in color.

[19] Although the quality of the photodiode's consciousness is the same quality generated by a binary thermistor, and many other simple mechanisms.

[20] Our ability to judge gradations in the level of consciousness when absolute levels are low may also be poor. As a loose metaphor, consider temperature. We are good at judging temperature as long as it fluctuates around the usual range, say between -50 and $+100$ °C. However, when temperature falls below that range, we become much less precise: both -200 and -273°C are inconceivably cold to us, and we certainly would not judge -200 to be much warmer than absolute zero. Similarly, a complex generating 1 or 10 bits of integrated information may feel a bit different (or rather 9 bits different), but it may feel like so little that, compared to our usual levels of consciousness, it essentially feels like nothing. Which is why, of course, it is good to have a thermometer or a Φ-meter.

[21] An optical metaphor can again be useful: things come crisply into existence at a certain focal distance, and with a certain exposure time. At shorter or longer focal distances things vanish out of focus: if exposure time is too short, they do not register; if it is too long, they blur.

[22] A photodiode or any other complex generating a quale consisting of just a single q-arrow.

[23] Here I ignore the issue of decomposing complex Boolean functions into elementary mechanisms.

[24] Note that each of these functions should be thought of as implemented according to its minimal formula (of shortest description length, *i.e.,* of minimal complexity). Clearly, minimal formulas that involve four inputs are more complex than formulas involving just one input (the parity function, for instance, is notoriously incompressible).

[25] While the particular combination of concepts described here was chosen for its familiarity (parity, symmetry, contiguousness, balance) rather than for informational efficiency, one can envision Boolean functions that realize "optimal" sets of concepts from the point of view of integrated information. For example, the four functions may be chosen so that, on average, the set of four output units jointly generate as much integrated information as possible, up to the theoretical maximum of 4 bits of Φ for every input string (by contrast, the "copy system," while transmitting all 4 bits in the input, would generate 4 times 1 bit of integrated information). Obviously, building a system that could respond optimally to a large set of input strings is exceedingly difficult (if at all possible), especially considering the need to build such a system using simple Boolean functions as building blocks.

[26] Again, it is difficult to build an optimal conceptual system that can preserve all the information in the input, corresponding in this case to 4 bits of integrated information for every input string.

[27] The extreme case is watching noisy "snow" patterns flickering on a TV screen. We treat the overwhelming majority of TV frames as equivalent, under the concept of "TV snow." If one were an optimal conceptual system, however, each frame would be conceptualized as its own very particular kind of pattern (say exhibiting a certain amount of 17th order symmetries, another amount of 11th order symmetries, belonging to the 6th class of contiguousness, *etc.*). In a sense, every noisy frame would be read as an astonishingly deep, rich, meaningful and unique pattern, perhaps as a work of art.

[28] Dreams prove that an adult brain does not need the outside world to generate experience "here and now": the mechanisms of the main complex within the brain are sufficient, all by themselves, to generate the informational relationships that constitute experience. Not to mention that in dreams we tend to be remarkably passive.

[29] Indeed, the shape of experience can be said to be the quintessential "phenotype."

Literature Cited

Albus, J. S., G. A. Bekey, J. H. Holland, N. G. Kanwisher, J. L. Krichmar, M. Mishkin, *et al.* 2007. A proposal for a Decade of the Mind initiative. *Science* **317:** 1321.

Alkire, M. T., A. G. Hudetz, and G. Tononi. 2008. Consciousness and anesthesia. *Science* **322:** 876–880.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press, New York.

Bachmann, T. 2000. *Microgenetic Approach to the Conscious Mind*. John Benjamins, Philadelphia.

Balduzzi, D., and G. Tononi. 2008. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* **4:** e1000091.

Bateson, G. 1972. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chandler, San Francisco.

Block, N., ed. 1978. *Trouble with Functionalism,* Vol. 9. Minnesota University Press, Minneapolis.

Blumenfeld, H., and J. Taylor. 2003. Why do seizures cause loss of consciousness? *Neuroscientist* **9:** 301–310.

Bower, J. M. 2002. The organization of cerebellar cortical circuitry revisited: implications for function. *Ann. N.Y. Acad. Sci.* **978:** 135–155.

Cover, T. M., and J. A. Thomas. 2006. *Elements of Information Theory,* 2nd ed. Wiley-Interscience, Hoboken, NJ.

Crick, F., and C. Koch. 2003. A framework for consciousness. *Nat. Neurosci.* **6:** 119–126.

Dehaene, S., C. Sergent, and J. P. Changeux. 2003. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl. Acad. Sci. USA* **100:** 8520–8525.

Dennett, D. C. 1991. *Consciousness Explained*. Little, Brown, Boston, MA.

Edelman, G. M. 1987. *Neural Darwinism: The Theory of Neuronal Group Selection*. BasicBooks, New York.

Feldman, J. 2003. A catalog of Boolean concepts. *J. Math. Psychol.* **47:** 75–89.

Gazzaniga, M. S. 2005. Forty-five years of split-brain research and still going strong. *Nat. Rev. Neurosci.* **6:** 653–659.

Hagmann, P., L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, *et al.* 2008. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6:** e159.

Hobson, J. A., E. F. Pace-Schott, and R. Stickgold. 2000. Dreaming and the brain: toward a cognitive neuroscience of conscious states. *Behav. Brain Sci.* **23:** 793–842.

Jackson, F. 1986. What Mary didn't know. *J. Philos.* **83:** 291–295.

Koch, C. 2004. *The Quest for Consciousness: A Neurobiological Approach*. Roberts, Denver, CO.

Koch, C., and G. Tononi. 2008. Can machines be conscious? *Spectrum IEEE* **45:** 55–59.

Koch, C., and N. Tsuchiya. 2007. Attention and consciousness: two distinct brain processes. *Trends Cogn. Sci.* **11:** 16–22.

Massimini, M., F. Ferrarelli, R. Huber, S. K. Esser, H. Singh, and G. Tononi. 2005. Breakdown of cortical effective connectivity during sleep. *Science* **309:** 2228–2232.

Massimini, M., F. Ferrarelli, S. K. Esser, B. A. Riedner, R. Huber, M. Murphy, *et al.* 2007. Triggering sleep slow waves by transcranial magnetic stimulation. *Proc. Natl. Acad. Sci. USA* **104:** 8496–8501.

Nørretranders, T. 1998. *The User Illusion: Cutting Consciousness Down to Size*. Viking, New York.

Palmer, S. E. 1999. Color, consciousness, and the isomorphism constraint. *Behav. Brain Sci.* **22:** 923–943; discussion 944–989.

Pashler, H. E. 1998. *The Psychology of Attention*. MIT Press, Cambridge, MA.

Posner, J. B., and F. Plum. 2007. *Plum and Posner's Diagnosis of Stupor and Coma,* 4th ed. Oxford University Press, New York.

Rovelli, C. 1996. Relational quantum mechanics. *Int. J. Theor. Phys.* **35:** 1637–1678.

Sporns, O., G. Tononi, and G. M. Edelman. 2000. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb. Cortex* **10:** 127–141.

Sporns, O., G. Tononi, and R. Kotter. 2005. The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* **1:** e42.

Steriade, M., I. Timofeev, and F. Grenier. 2001. Natural waking and sleep states: a view from inside neocortical neurons. *J. Neurophysiol.* **85:** 1969–1985.

Tononi, G. 2001. Information measures for conscious experience. *Arch. Ital. Biol.* **139:** 367–371.

Tononi, G. 2004. An information integration theory of consciousness. *BMC Neurosci.* **5:** 42.

Tononi, G., and G. M. Edelman. 1998. Consciousness and complexity. *Science* **282:** 1846–1851.

Tononi, G., and S. Laureys. 2008. The neurology of consciousness: an overview. Pp. 375–412 in *The Neurology of Consciousness*, S. Laureys and G. Tononi, eds. Elsevier, Oxford.

Tononi, G., and O. Sporns. 2003. Measuring information integration. *BMC Neurosci.* **4:** 31.

Tononi, G., O. Sporns, and G. M. Edelman. 1996. A complexity measure for selective matching of signals by the brain. *Proc. Natl. Acad Sci. USA* **93:** 3422–3427.

van Zandvoort, M. J., T. C. Nijboer, and E. de Haan. 2007. Developmental colour agnosia. *Cortex* **43:** 750–757.

Wheeler, J. A., and K. W. Ford. 1998. *Geons, Black Holes, and Quantum Foam: A Life in Physics*, 1st ed. Norton, New York.