Integrated information theory: From consciousness to its physical substrate

Article *in* Nature Reviews Neuroscience · May 2016

DOI: 10.1038/nrn.2016.44

CITATIONS
462

READS
10,050

4 authors:

Giulio Tononi
University of Wisconsin–Madison
577 PUBLICATIONS **50,029** CITATIONS
SEE PROFILE

Melanie Boly
University of Wisconsin–Madison
275 PUBLICATIONS **19,402** CITATIONS
SEE PROFILE

Marcello Massimini
University of Milan
172 PUBLICATIONS **12,924** CITATIONS
SEE PROFILE

Christof Koch
Allen Institute for Brain Science
908 PUBLICATIONS **85,359** CITATIONS
SEE PROFILE

OPINION

Integrated information theory: from consciousness to its physical substrate

Giulio Tononi, Melanie Boly, Marcello Massimini and Christof Koch

Abstract | In this Opinion article, we discuss how integrated information theory accounts for several aspects of the relationship between consciousness and the brain. Integrated information theory starts from the essential properties of phenomenal experience, from which it derives the requirements for the physical substrate of consciousness. It argues that the physical substrate of consciousness must be a maximum of intrinsic cause–effect power and provides a means to determine, in principle, the quality and quantity of experience. The theory leads to some counterintuitive predictions and can be used to develop new tools for assessing consciousness in non-communicative patients.

Consciousness is subjective experience — 'what it is like', for example, to perceive a scene, to endure pain, to entertain a thought or to reflect on the experience itself[1–3]. When consciousness fades, as it does in dreamless sleep, from the intrinsic perspective of the experiencing subject, the entire world vanishes.

Consciousness depends on the integrity of certain brain regions and the particular content of an experience depends on the activity of neurons in parts of the cerebral cortex[4]. However, despite increasingly refined clinical and experimental studies, a proper understanding of the relationship between consciousness and the brain has yet to be established[5,6]. For example, it is not known why the cortex supports consciousness when the cerebellum does not, despite having four times as many neurons[7,8], or why consciousness fades during deep sleep while the cerebral cortex remains active. There are also many other difficult questions about consciousness. Are patients with a functional island of cortex surrounded by widespread damage conscious, and if so, of what? Are newborn infants conscious? Are animals that display complex behaviours, but have brains very different from humans, conscious[6]? Can intelligent machines be conscious[9]?

To answer these questions, the empirical study of consciousness should be complemented by a theoretical approach. The reason why some neural mechanisms, but not others, should be associated with consciousness has been called 'the hard problem' because it seems to defy the possibility of a scientific explanation[10]. In this Opinion article, we provide an overview of the integrated information theory (IIT) of consciousness, which has been developed over the past few years[1–3,11,12]. IIT addresses the hard problem in a new way. It does not start from the brain and ask how it could give rise to experience; instead, it starts from the essential phenomenal properties of experience, or axioms, and infers postulates about the characteristics that are required of its physical substrate. Moreover, IIT presents a mathematical framework for evaluating the quality and quantity of consciousness[1–3,9]. We begin by providing a summary of the axioms and corresponding postulates of IIT and show how they can be used, in principle, to identify the physical substrate of consciousness (PSC). We then discuss how IIT explains in a parsimonious manner a variety of facts about the relationship between consciousness and the brain, leads to testable predictions, and allows inferences and extrapolations about consciousness.

From phenomenology to physics

The axioms of IIT state that every experience exists intrinsically and is structured, specific, unitary and definite. IIT then postulates that, for each essential property of experience, there must be a corresponding causal property of the PSC. The postulates of IIT state that the PSC must have intrinsic cause–effect power; its parts must also have cause–effect power within the PSC and they must specify a cause–effect structure that is specific, unitary and definite. Below, we discuss the axioms and postulates of IIT (see Supplementary information S1,S2 (figure, box)) and describe the fundamental identity — between an experience and a conceptual structure — that it proposes (FIG. 1).

The first axiom of IIT states that experience exists intrinsically. As recognized by Descartes[13], my own experience is the only thing whose existence is immediately and absolutely evident, and it exists for myself, from my own intrinsic perspective. The corresponding postulate states that the PSC must also exist intrinsically. For something to exist in a physical sense, it must have cause–effect power — that is, it must be possible to make a difference to it (that is, change its state) and it must be able to make a difference to something. Moreover, the PSC must exist intrinsically — that is, it must have cause–effect power for itself, from its own intrinsic perspective. A neuron in the brain, for example, satisfies the criterion for existence because it has two or more internal states (such as active and inactive) that can be affected by inputs (causes) and its output can make a difference to other neurons (effects). A minimal system consisting of two interconnected neurons satisfies the criterion of intrinsic existence because, through their reciprocal interactions, the system can make a difference to itself.

The axiom of composition states that experience is structured, being composed of several phenomenal distinctions that exist within it. For example, within an experience, I may distinguish a piano, a blue colour, a book, countless spatial locations, and so on

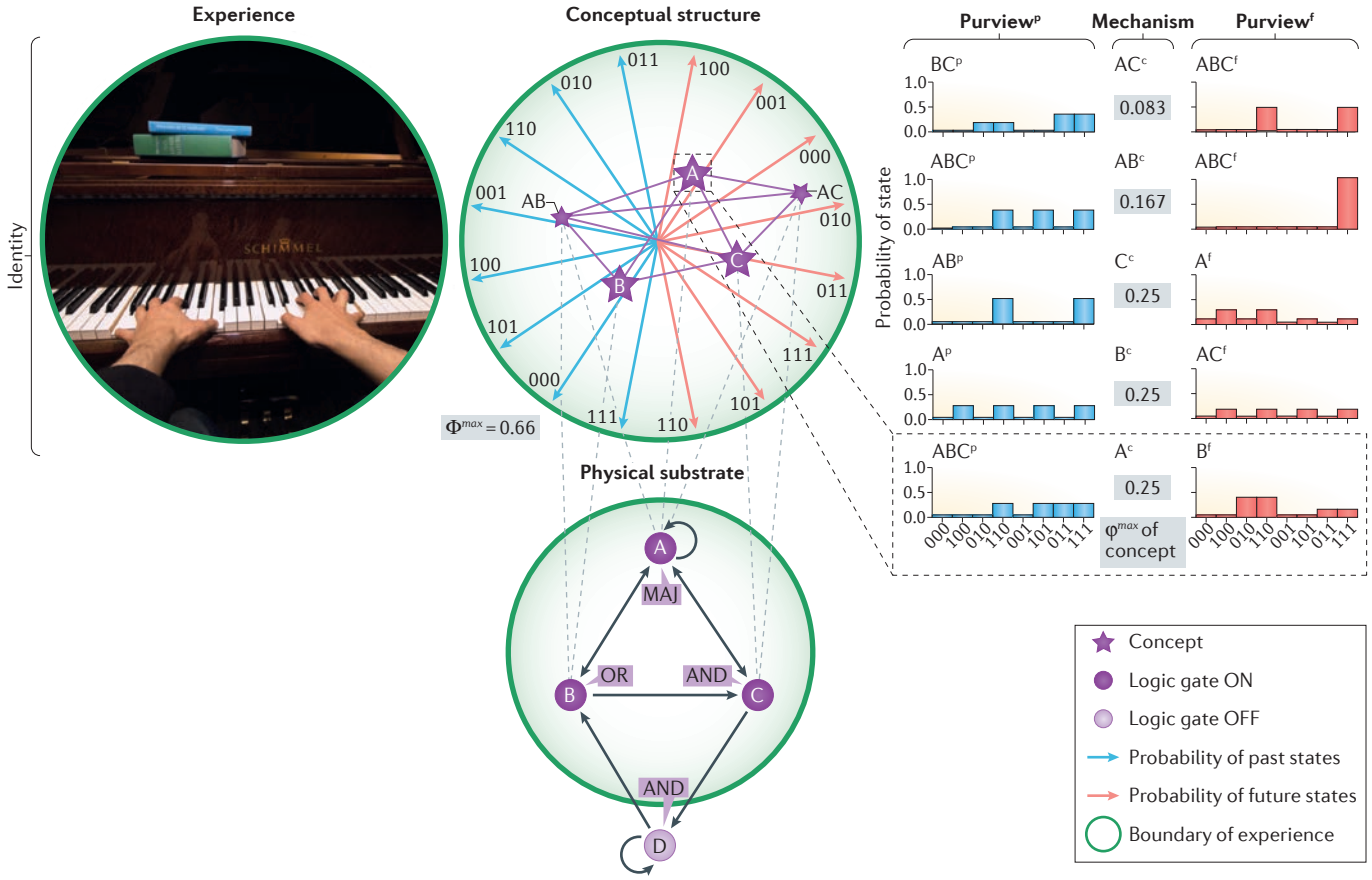


Figure 1 | **An experience is a conceptual structure.** According to integrated information theory (IIT), a particular experience (illustrated here from the point of view of the subject) is identical to a conceptual structure specified by a physical substrate. The true physical substrate of the depicted experience (seeing one's hands on the piano) and the associated conceptual structure are highly complex. To allow a complete analysis of conceptual structures, the physical substrate illustrated here was chosen to be extremely simple[1,2]: four logic gates (labelled A, B, C and D, where A is a Majority (MAJ) gate, B is an OR gate, and C and D are AND gates; the straight arrows indicate connections among the logic gates, the curved arrows indicate self-connections) are shown in a particular state (ON or OFF). The analysis of this system, performed according to the postulates of IIT, identifies a conceptual structure supported by a complex constituted of the elements A, B and C in their current ON states. The borders of the complex, which include elements A, B, and C but exclude element D, are indicated by the green circle. According to IIT, such a complex would be a physical substrate of consciousness (Supplementary information S1 (figure)). The conceptual structure is represented as a set of stars and, equivalently, as a set of histograms. The green circle represents the fact that experience is definite (it has borders). Each histogram illustrates the cause–effect repertoire of a concept: how a particular mechanism constrains the probability of past and future states of its maximally irreducible purview within the complex ABC. The bins on the horizontal axis at the bottom of the histograms represent the 16-dimensional cause–effect space of the complex — all its eight possible past states (p; in blue) and eight possible future states (f; in red; ON is 1 and OFF is 0). The vertical axis represents the probability of each state (for consistency, the probability values shown are over the states of the entire complex and not just over the subset of elements constituting the purview). In this example, five of seven possible concepts exist, specified by the mechanisms A, B, C, AB, AC (all with $\varphi^{max}>0$) in their current state (which are labelled as A^c, B^c, etc.). The subsets BC and ABC do not specify any concept because their cause–effect repertoire is reducible by partitions ($\varphi^{max}=0$). In the middle, the 16-dimensional cause–effect space of the complex is represented as a circle, where each of the 16 axes corresponds to one of the eight possible past (p; blue arrows) and eight possible future states (f; red arrows) of the complex, and the position along the axis represents the probability of that state. Each concept is depicted as a star, the position of which in cause–effect space represents how the concept specifies the probability of past and future states of the complex, and the size of which measures how irreducible the concept is (φ^{max}). Relations between two concepts (overlaps in their purviews) are represented as lines between the stars. The fundamental identity postulated by IIT claims that the set of concepts and their relations that compose the conceptual structure are identical to the quality of the experience. This is how the experience feels — what it is like to be the complex ABC in its current state 111. The intrinsic irreducibility of the entire conceptual structure (Φ^{max}, a non-negative number) reflects how much consciousness there is (the quantity of the experience). The irreducibility of each concept (φ^{max}) reflects how much each phenomenal distinction exists within the experience. Different experiences correspond to different conceptual structures.

(FIG. 1). Based on this axiom, IIT postulates that the elements that constitute the PSC must also have cause–effect power within the PSC, either alone or in combination (composing first-order and higher-order mechanisms, respectively).

The axiom of information states that experience is specific, being composed of a particular set of phenomenal distinctions (qualia), which make it what it is and different from other experiences. In the example shown in FIG. 1, the content of my current experience might be composed of seeing a book (rather than seeing no book), which is blue (rather than not blue), and so on for all other possible contents of consciousness. The corresponding postulate states that the PSC must specify a cause–effect structure

of a specific form, which makes it different from other possible forms. A cause–effect structure is defined as the set of cause–effect repertoires specified by all the mechanisms of a system. A cause–effect repertoire specifies how a mechanism in its current state affects the probability distribution of past and future states of the system.

The axiom of integration states that experience is unitary, meaning that it is composed of a set of phenomenal distinctions, bound together in various ways, that is irreducible to non-interdependent subsets. For example, I experience a whole visual scene and that experience cannot be subdivided into independent experiences of the left and right sides of the visual field. In other words, the content of an experience (information) is integrated within a unitary consciousness. The corresponding postulate states that the cause–effect structure specified by the PSC must also be unitary — that is, it must be irreducible to the cause–effect structure specified by non-interdependent subsystems. Note that, from the intrinsic perspective of the system, integration requires that every part of the system has both causes and effects within the rest of the system, which implies bidirectional interactions. The irreducibility of a conceptual structure is measured as integrated information (denoted Φ, the minimum distance between an intact and a partitioned cause–effect structure). The integration postulate also requires the irreducibility of each cause–effect repertoire (denoted φ, the minimum distance between an intact and a partitioned cause–effect repertoire) and the irreducibility of relations among overlapping cause–effect repertoires.

The axiom of exclusion states that an experience is definite in its content and spatio-temporal grain. For example, in the scene depicted in FIG. 1, the content of my present experience includes seeing my hands on the piano, the books on the piano, one of which is blue, and so on, but I am not having an experience with less content (for example, the same scene in black and white, lacking the phenomenal distinction between coloured and not coloured) or with more content (for example, including the additional phenomenal distinction of feeling one's blood pressure as high or low). The duration of the instant of consciousness is also definite, ranging from a few tens of milliseconds to a few hundred milliseconds, rather than lasting a few microseconds or a few minutes[14–16]. The corresponding postulate states that the cause–effect structure specified by the PSC must also be definite. It must specify a definite set of cause–effect repertoires over a definite set of elements, neither less nor more, at a definite spatio-temporal grain, neither finer nor coarser. Because a prerequisite for intrinsic existence is having irreducible cause–effect power, the cause–effect structure that actually exists, over a set of elements and spatio-temporal grains, is that which is maximally irreducible (Φ^{max}), called a conceptual structure. As a consequence, any cause–effect structure overlapping over the same set of elements and spatio-temporal grain is excluded. The exclusion postulate also requires the maximum irreducibility of cause–effect repertoires (denoted φ^{max}), called concepts, and of relations among overlapping concepts.

A set of elements in a state that satisfies all the postulates of IIT constitutes the PSC and is referred to as a complex (FIG. 1). Thus a complex specifies a conceptual structure composed of concepts, which can be represented as a set of points (shown as a constellation of stars in FIG. 1) in cause–effect space, in which each axis corresponds to a possible past and future state of the system and each star corresponds to a concept[1] (FIG. 1). With these notions at hand, the fundamental identity of IIT can be stated as follows[2]: an experience is identical to a conceptual structure, meaning that every property of the experience must correspond to a property of the conceptual structure and vice versa. Note that the postulated identity is between an experience and the conceptual

Glossary

Achromatopsia
A condition in which a person is unable to perceive colours.

Anosognosia
A condition in which a person has a neurological deficit, but is unaware of it.

Axioms
Properties that are self-evident and essential; in integrated information theory, those that are true of every possible experience — namely, intrinsic existence, composition, information, integration and exclusion.

Background conditions
Factors that enable consciousness, such as neuromodulators and external inputs that maintain adequate excitability.

Cause–effect repertoire
The probability distribution of potential past and future states of a system that is specified by a mechanism in its current state.

Cause–effect space
A space with each axis representing the probability of each possible past and future state of a system.

Cause–effect structure
The set of cause–effect repertoires specified by all the mechanisms of a system in its current state.

Complex
A set of elements in a state that specifies a conceptual structure corresponding to a maximum of integrated information (Φ^{max}). A complex is thus a physical substrate of consciousness.

Concepts
The cause–effect repertoires specified by a mechanism that is maximally irreducible (φ^{max}).

Conceptual structure
The set of all concepts specified by a system of elements in a state with their respective φ^{max} values, which can be plotted as a set of points in cause–effect space.

Content-specific NCC
Neural elements, the activity of which determines a particular content of experience.

Elements
The minimum constituents of a system that have at least two different states (for example, being on or off), inputs that can affect those states and outputs that depend on them.

Full NCC
The neural elements constituting the physical substrate of consciousness, irrespective of its specific content.

Integrated information
(Denoted Φ). Information that is specified by a system that is irreducible to that specified by its parts. It is calculated as the distance between the conceptual structure specified by the intact system and that specified by its minimum information partition.

Mechanism
Any subset of elements within a system that has cause–effect power on it (that is, that constrains its cause–effect space).

Neural correlates of consciousness
(NCC). The minimum neuronal mechanisms jointly sufficient for any one specific conscious experience.

Postulates
Properties of experience that are derived from the axioms of integrated information theory and that must be satisfied by the physical substrate of consciousness — namely, to be a maximum of irreducible, specific, compositional, intrinsic cause–effect power (intrinsic cause–effect power for short).

Purviews
The subsets of elements of a complex, the past and future states of which are constrained by a mechanism specifying a concept.

Qualia
The qualitative feeling of phenomenal distinctions within an experience (for example, seeing a colour, hearing a sound or feeling a pain).

Relations
Maximally irreducible overlaps among the purviews of two or more concepts.

structure specified by the PSC, not between an experience and the set of elements in a state constituting the PSC (FIG. 1). The quality or content of consciousness — which particular way the system exists for itself — corresponds to the form of the conceptual structure. The quantity of consciousness — how much the system exists for itself — corresponds to its irreducibility Φ^{max}.

The PSC within the brain

Experimental evidence currently suggests that the neural correlates of consciousness (NCC) are likely to be located in certain parts of the cortico-thalamic system[5], but it is not known specifically which cortical areas, layers or neuronal populations are involved, whether the relevant units are neurons or groups of neurons, and which aspects of their activity matter[5]. It is also not known whether the neural substrate of consciousness is anatomically fixed or can shrink, expand and move. IIT offers theoretical clarity on the empirical notion of the NCC[5]. Specifically, it states that the content-specific NCC correspond to the neural elements of the PSC in a particular state (activity pattern), which specify a particular phenomenal content; the full NCC correspond to the neural elements constituting the PSC irrespective of their particular state; the background conditions are factors that enable consciousness, such as neuromodulators and external inputs that maintain adequate excitability, which are kept fixed when evaluating the Φ value of the PSC. Most importantly, the axioms and postulates of IIT can be used to provide a single, general principle for identifying the PSC in the brain — namely that the PSC must correspond to a complex of neural elements with maximum intrinsic cause–effect power.

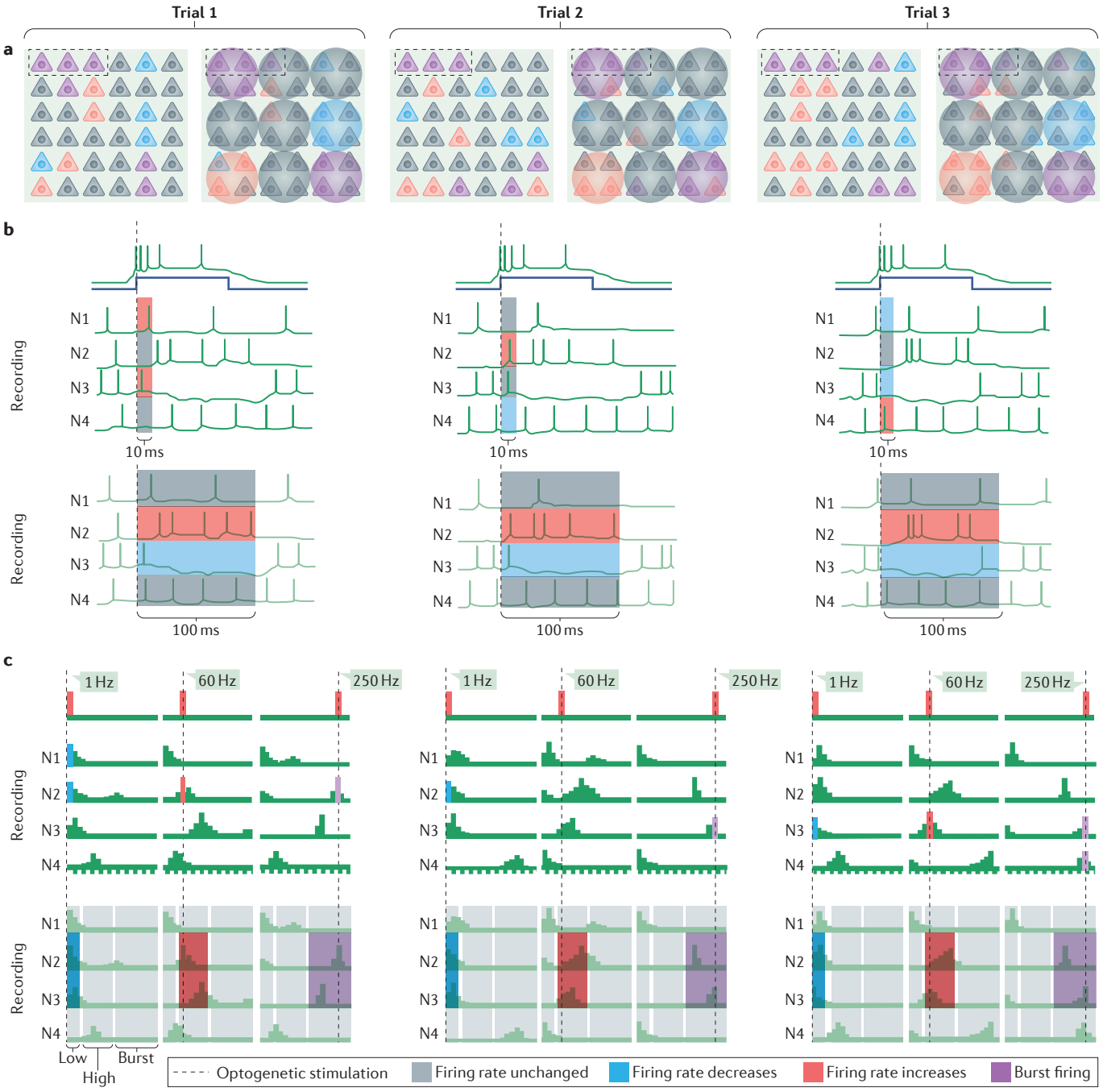
Elements of the PSC. What is the spatial scale of the neural elements that support consciousness: synapses, neurons, neuronal groups, local fields or perhaps all of these? According to IIT, the neural elements of the PSC are those, and only those, that support a maximum of cause–effect power, as determined from the intrinsic perspective of the system itself. Importantly, and contrary to common reductionist assumptions[17], cause–effect power can be higher at a macro-level than at a micro-level[18]. For example, a system of neuron-like micro-elements may have less cause–effect power than the same system coarse-grained at the macro-level of neuronal groups (FIG. 2a). In general, whether the macro or micro grain size has higher cause–effect power depends on how intra- and inter-group connections are organized and the amount of indeterminism (noise) and degeneracy (multiple ways of obtaining the same effect[18]).

An exhaustive evaluation of cause–effect power at multiple levels is only possible in small simulated networks[19]. In a real network[20], we could start by assessing the cause–effect repertoire of individual neurons. For example, if a neuron is firing a burst of spikes, its cause repertoire is the probability distribution of past network states that would have caused it to burst (for example, firing patterns of its afferent neurons within the previous 100 ms). Similarly, its effect repertoire is the probability distribution of future network states given that the neuron is bursting. Experimentally, we could obtain an estimate of such cause–effect repertoires by stimulating one or more neurons optogenetically while simultaneously recording the firing activity of a population of neurons via two-photon calcium imaging (keeping the background conditions constant, such as the level of arousal and sensory input) (FIG. 2a). Next, we would need to test for the irreducibility of the cause–effect repertoires, which can be achieved by noising connections (that is, enforcing firing at chance levels) across a partition of the network. Doing so would establish which subset of incoming connections makes the most irreducible difference (φ^{max}) to the firing of the observed neuron[1] (and this could be carried out analogously for outgoing connections). A similar procedure should then be repeated for subsets of two neurons, three neurons, and so on, because combinations of neurons can also have irreducible cause–effect repertoires (defined as higher order mechanisms). Such experiments would provide an estimate of maximally irreducible cause–effect repertoires at the level of neurons.

To evaluate cause–effect power at the macro-level, we could then repeat the same stimulation–recording–noising procedure by considering subsets of neurons as distinct macro-groups and mapping micro-states onto macro-states. For example, we could take all pyramidal neurons in each mini-column as a distinct group and define the group state as low firing, high firing or bursting, depending on the overall firing rate of the neurons over 100 ms. By estimating the φ^{max} value of cause–effect repertoires at the level of both individual neurons and groups of neurons, an experimenter could thus assess at which grain size the network has most cause–effect power from its own intrinsic perspective — that is, at which level it makes the most difference to itself. IIT predicts that the elements of the PSC are to be found at exactly that level and not at any finer or coarser grain, a prediction that is empirically testable: does the firing of a single neuron make a difference[21] to the content of experience, or only the average activity of a cortical mini-column[22]?

Timescale. Which timescale of neuronal activity is important for consciousness: a few milliseconds, tens of milliseconds, hundreds of milliseconds, or perhaps all of these? Again, IIT predicts that the relevant time interval should be that which makes the most difference to the system, as determined from its intrinsic perspective. Once more, depending on the specific mechanisms of a system, some macro-temporal grain may have a higher cause–effect power than both finer and coarser grains (FIG. 2b). Whatever timescale turns out to have the maximum cause–effect power within the relevant brain regions, it should be consistent with estimates of the timescale of experience[14–16].

State of the elements. An external observer can choose to analyse brain states at any level of detail. For example, some neurophysiologists may be interested in the effects of the timing of individual neuronal spikes on brain function, others in the effects of broad fluctuations in the activity of populations of neurons. In fact, it is likely that almost any change in the state of any neurobiological variable will have some effect somewhere in the brain[21]. According to IIT, the neural states that are important for consciousness are only those that have maximum cause–effect power on the system itself. For example, assume that, from the intrinsic perspective of the system, maximum cause–effect power was achieved when coarse-graining firing states into low, high and burst firing (FIG. 2c). In this case, IIT predicts that finer grained neural states, despite their demonstrable neurophysiological effects, make no difference to the content of experience. Note that spatio-temporal grain and the relevant activity states of the elements specifying the PSC could change according to brain region, developmental period, species, neuromodulatory milieu and even the task being performed.



Constitution of the PSC. Assume that we have determined that the elementary units of the PSC are local groups of cortical neurons, over a time interval of ~100 ms, with three relevant states (low, high and burst firing) (FIG. 3a). Next we must determine, at the system level, which particular subset of neuronal groups constitutes the PSC for a particular experience. IIT addresses this question from first principles — it predicts that the PSC is the set of neuronal groups that has maximally irreducible cause–effect power on itself, specifying a conceptual structure with the highest value of Φ^1 (FIG. 3b). Ideally, systematic manipulation and recording of this particular set of neuronal groups would show that it has the maximum value of Φ, whereas any other assortment of neuronal groups in the brain has a lower value of Φ.

Although such an exhaustive evaluation of Φ is not currently feasible, neuroimaging studies can evaluate two key requirements for a high Φ value: information, using measures that reflect the size of the repertoire of neural states the system can have (that is, neurophysiological differentiation)[23]; and integration, using measures of functional or effective connectivity among brain regions[24,25]. In addition, large-scale computer simulations based on the known anatomy and physiology of cortical circuits[26] can be used to assess cause–effect repertoires, test their irreducibility and estimate conceptual structures. Crucially, if the evidence thus obtained indicates that the PSC does not correspond to a maximum of intrinsic cause–effect power, IIT would be invalidated. A related prediction is

◀ Figure 2 | **Identifying the elements, timescale and states of the physical substrate of consciousness (PSC) from first principles.** It is possible to determine maxima of cause–effect power within the central nervous system by perturbing and observing neural elements at various micro- and macro-levels[18]. High cause–effect power is reflected in deterministic responses and low cause–effect power is reflected in responses that vary randomly across trials. **a** | To identify the spatial grain of the elements of the PSC supporting consciousness, a schematic example shows how optogenetic perturbation and unit recording could be applied to a subset of neurons (here, 3 out of 36 neurons) to establish maxima of cause–effect power. For each of three trials, the left panel shows the effects of the perturbation on the entire system at the micro-level. Grey neurons are unaffected, blue neurons decrease their firing rates, red neurons increase their firing rates and purple neurons respond with burst firing. The right-hand panel shows the effects of the perturbation at the macro-level after coarse-graining of the 36 neurons into nine groups of four cells each. Macro-states are defined according to the rule that if ≥50% of the neurons in the group are in a given micro-state (such as low firing, high firing or bursting), then the group is considered to be in that state at the macro-level. In this example, the macro-level (groups of neurons) has higher cause–effect power than the micro-level (single neurons), because the response is deterministic at the macro-level (as evidenced by the consistent colour scheme), whereas there are variations between trials at the micro-level (inconsistent colours). **b** | To identify the temporal grain of neuronal activity supporting consciousness, a possible experimental setup would be one in which one neuron (the top trace) is optogenetically excited while recording from other neurons (labelled N1–N4) across three trials, shown in the upper panel at the 10 ms timescale (micro-scale). Grey shading indicates no effects on neuron firing in the 10 ms following the stimulation compared with the 10 ms before the stimulation, blue shading indicates decreased firing and red shading indicates increased firing. The lower panel shows the same data after temporal coarse-graining over 100 ms intervals. Macro-states are defined according to the rule that if a neuron increases (or decreases) its firing rate by >50% within 100 ms post-stimulus compared with the baseline, the macro state is considered to be high (or low) firing. In this example, the macro-level (100 ms intervals) has higher cause–effect power (more deterministic responses) than the micro-level (10 ms intervals). **c** | To identify the neural states that support consciousness, optogenetic perturbations could be used to drive one neuron to fire either at low frequency, high tonic frequency or bursting (top trace) resulting in spectral peaks at 2 Hz (green), 50 Hz (red) and 150 Hz (yellow) for neurons N1–N4 (data are shown as a firing rate histogram). For each trial, the upper panel shows the responses of the other four neurons to each stimulation frequency at the micro-scale level in the spectral domain (micro-bins, only a few of which are represented). The coloured bars indicate coincidence, within a micro-bin, between the frequency of stimulation and the spectral peak of the responses. The lower panel of each trial shows the effect of the perturbation at the corresponding macro-level after spectral coarse-graining. Macro-states map into micro-states as indicated below the frequency bins. Here, spectral coarse-graining (binning firing rates into three levels, low, high and burst firing) results in higher cause–effect power (responses that are more deterministic) than at the micro-level.

that any perturbation of the PSC at the appropriate spatio-temporal grain should produce a change in experience, whereas any perturbation that does not alter the PSC should not.

Can the PSC change? An important issue is the extent to which the set of neural elements that constitute the PSC is fixed. Clearly, if a cortical area is inactivated (by a lesion, for example) it will no longer be part of the PSC and the phenomenal distinctions contributed by that area will no longer be available. For example, if cortical areas responding to colour are inactivated (FIG. 3c), experiences will not only lack colour, but patients would not even understand what is lacking (as reported in cases of achromatopsia with anosognosia[27]).

It is an open question whether the PSC can shrink, expand or move during normal wakefulness, possibly through attentional modulation of excitability and functional connectivity. For example, when we are engrossed in an action movie and not engaged in self-reflection, the activity in prefrontal areas decreases[28]. Does this mean that the PSC shrinks, like when colour areas are inactivated, or that brain regions supporting self-reflection remain inside the PSC but are inactive, in the same way that colour areas are inactive when watching a black and white movie? The location and size of the PSC is likely to change during sleep, during seizures, in patients with conversion and dissociative disorders, and possibly during hypnosis. During slow wave sleep, for example, neurons are bistable and show off-periods during which they become hyperpolarized (down-states) and silent[29]. However, these off-periods are usually not global, but affect local subsets of brain areas at different times[30]. Hence it is possible that during slow wave sleep the PSC may become smaller and reconfigure substantially. Sustained inactivation of certain areas during sleep may make dreaming patients incapable of reflective thought. Similarly, experiences of pure thought that have minimal perceptual content may be caused by slow waves that inactivate the posterior cortex, and be specified by a PSC that is considerably different from the PSC for purely perceptual experiences[31] (FIG. 3d). At other times, transient, local slow waves (indicative of an off-period) in colour areas may cause the PSC to shrink and lead to brief episodes of achromatopsia. Novel methods that allow the transient inactivation of specific cortical areas in humans, such as transcranial magnetic stimulation or focused ultrasound, would be ideal for evaluating the contribution of those areas to conscious content.

Multiple complexes. According to IIT, two or more non-overlapping complexes may coexist as discrete PSCs within a single brain[1], each with its own definite borders and value of Φ^{max}. The complex that specifies a person's day-to-day stream of consciousness should have the highest value of Φ^{max} — that is, it should be the 'major' complex. In some conditions, for example after a split-brain operation, the major complex may split (FIG. 3e). In such instances, one consciousness, supported by a complex in the dominant hemisphere and with privileged access to Broca's area, would be able to speak about the experience, but would remain unaware of the presence of another consciousness, supported by a complex in the other hemisphere, which can be revealed by carefully designed experiments[32,33]. An intriguing possibility is that splitting of the PSC may also occur in healthy people during long-lasting dual-task conditions — for example, when driving in an auto-pilot like manner on a familiar road while listening to an engaging conversation (FIG. 3f). Splitting into separate maxima may also occur through functional disconnections caused by pathological conditions, such as conversion and dissociative disorders[34].

Another intriguing possibility is that multiple conscious streams may coexist within a single brain in daily life. For example, the grid-like architectures in the colliculus and related mesencephalic regions, which are adept at multimodal integration within a spatial framework, may support a separate minor complex. Some examples of high-level cognitive performance such as judging whether a scene is congruous or incongruous[35,36] — that appear to be carried out unconsciously from the perspective of the major complex — may support a separate minor complex (FIG. 3e,g).

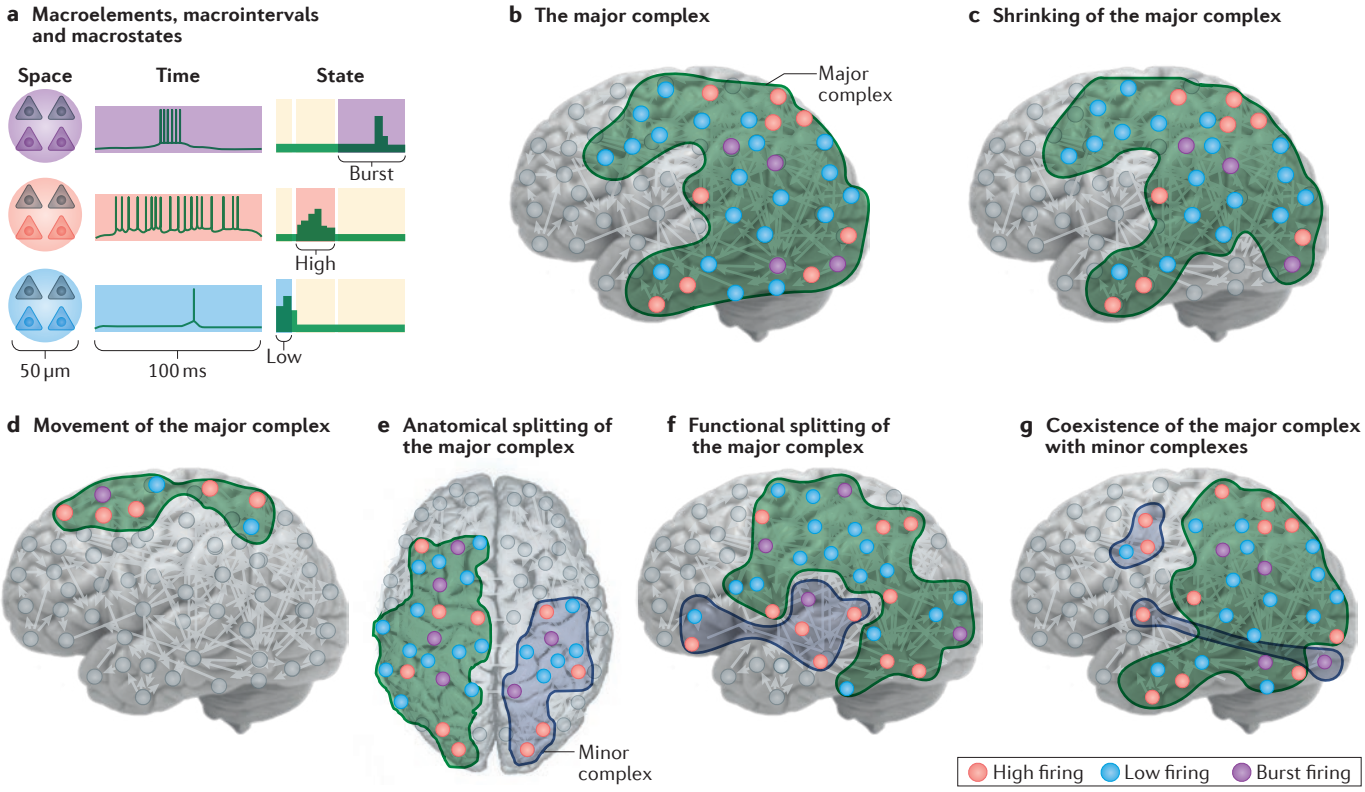


Figure 3 | **Identifying the physical substrate of consciousness (PSC) from first principles.** The complex of neural elements that constitutes the PSC can be identified by searching for maxima of intrinsic cause–effect power. **a** | For example, assume that the elements, timescale and states at which intrinsic cause–effect power reaches a maximum have been identified using optogenetic and unit recording tools (FIG. 2). Here, the elements are groups of neurons, the timescale is over 100 ms and there are three states (low, high and burst firing). **b** | In a healthy, awake participant, the set of neural elements specifying the conceptual structure with the highest Φ^{max} is assumed, based on current evidence, to be a complex of neuronal groups distributed over the posterior cortex and portions of the anterior cortex[5]. Empirical studies can, in principle, establish whether the full neural correlates of consciousness[5] correspond to the maximum of intrinsic cause–effect power, thereby corroborating or falsifying a key prediction of integrated information theory. **c** | The boundaries of the PSC (green line) may change after cortical lesions, such as those causing absolute achromatopsia, resulting in a smaller PSC. **d** | The PSC boundaries may also move as a result of changes in excitability and effective connectivity, as might occur during pure thought that is devoid of sensory content. **e** | The PSC could also split into two large local maxima of cause–effect power (represented here by green and blue boundaries) as a result of anatomical disconnections, such as in split-brain patients, in which instance each hemisphere would have its own consciousness. **f** | The PSC may also split as a result of functional disconnections, which may occur in some psychiatric disorders and perhaps under certain dual-task conditions — for example while driving and talking at the same time. **g** | The coexistence of a large major complex with one or more minor complexes that may support sophisticated, seemingly unconscious performance could be a common occurrence in everyday life.

Alternatively, some of these functions may be mediated by feedforward circuits[37] that have $\Phi^{max}=0$ because they lack integration and therefore are strictly unconscious[1]. An important question for the future is whether automatic, unconscious behaviours are mediated by specific cell types within the cortex, such as subcortical projection neurons of layer 5B[38], that are different from other cell types that support consciousness.

Information capacity of consciousness

The information-processing approach common in psychology estimates the information capacity of human consciousness to be at around 7 ± 2 items[39] or ≤40 bits per second[39,40]. In the classic Sperling task[41], for example, participants are presented with a set of 12 letters for 300 ms, of which, after a mask and a delay, they can report at most four (FIG. 4). The inference from such experiments is that the information content of consciousness is extremely limited, as is also suggested by the attentional blink and related psychophysical paradigms[42,43]. For example, in change blindness, a major modification in a visual scene may go undetected if a blank is interposed between the two images[44]. In this view, the content of consciousness is limited to what can be accessed and reported, despite our phenomenal impression of richer content[42,45,46]. By contrast, others argue that phenomenal consciousness (what it is like to have an experience) has far greater capacity than access consciousness (what can be reported)[47–49]. For example, if participants are cued to a particular row of the Sperling display during the delay period, they can report three letters of any row; moreover, they can report the colour diversity of unattended letters at no cost to the identification of the cued letters[50]. Likewise, change blindness may be due not to a failure to experience, but to a lack of memory for the experience[51]. Similarly, low-level phenomenal features may be difficult to report because they vary rapidly and may be forgotten before they can be accessed from top-down mechanisms; pre-categorical stimuli, such as irregular scribbles, may be phenomenally salient but hard to describe in words.

IIT claims that human consciousness has a high capacity for integrated information (BOX 1). Even for a simple experience, such as seeing the Sperling display, the elements

Box 1 | **Consciousness, integrated information and Shannon information**

The term information is used very differently in integrated information theory (IIT) and in Shannon's theory of communication[1], and confusing the two meanings can cause misunderstandings[80]. The word information derives from the Latin verb *informare*, which means 'to give form'. In IIT the information content of an experience is specified by the form of the associated conceptual structure (the quality of the integrated information) and quantified by Φ^{max} (the quantity of integrated information). In IIT, information is causal and intrinsic: it is assessed from the intrinsic perspective of a system based on how its mechanisms and present state affect the probability of its own past and future states (cause–effect power). It is also compositional, in that different combinations of elements can simultaneously specify different probability distributions within the system. Moreover, it is qualitative, as it determines not only how much a system of mechanisms in a state constrains its past and future states, but also how it does so. Crucially, in IIT, information must be integrated. This means that if partitioning a system makes no difference to it, there is no system to begin with. Information in IIT is exclusive — only the maxima of integrated information are considered. By contrast, Shannon information is observational and extrinsic — it is assessed from the extrinsic perspective of an observer and it quantifies how accurately input signals can be decoded from the output signals transmitted across a noisy channel. It is not compositional nor qualitative, and it does not require integration or exclusion[1].

When averaged over many different states of the physical substrate of consciousness (PSC), we can think of the integrated information Φ^{max} as a measure of the intrinsic phenomenal capacity of the conceptual structures specified by the PSC. By contrast, Shannon information can be used to measure the extrinsic access capacity of a channel that runs from a subset of elements of the PSC to Broca's area and from there to the motor neurons that ultimately convey the report (FIG. 4). In IIT, the experience of seeing the Sperling display is identical to a particular conceptual structure — it is a form in cause–effect space with a high value of integrated information Φ^{max}, as specified by its PSC (FIG. 4). The average value of Φ^{max} for different states of the PSC measures its intrinsic phenomenal capacity. The figure also shows a neural information channel from the PSC to Broca's area, formed dynamically by top-down attentional mechanisms located in the prefrontal cortex, which select which subset of elements of the PSC should drive the report (FIG. 4). This channel conveys extrinsic information and has a low Shannon capacity (only four letters at a time can be reported), which corresponds to the mutual information between its inputs and outputs. Seen in this way, it becomes obvious that the extrinsic information that can be selected through attention, kept in working memory and channelled out for report is only a partial read-out of the intrinsic information that is specified by the PSC over its own cause–effect space. Although at any given time we can access and report the state of a few elements of the PSC, and that of some other elements at another time, it is not possible to dump the state of all elements through a limited capacity channel. It is certainly not possible to transmit a conceptual structure (intrinsic information) through a channel (extrinsic information)—phenomenal capacity, properly understood, truly exceeds access capacity. Likewise, conscious information is not something that is transmitted or broadcast from one part of the brain to another[77,78] (Supplementary information S5 (box)).

of the PSC specify a rich conceptual structure (high Φ^{max}) composed of a very large number of concepts and relations. These correspond to all the phenomenal distinctions that make that experience what it is and thereby different from countless others[11] (FIG. 4). It is useful to distinguish between low- and high-order concepts, depending on how many PSC elements are contained in their purviews. For example, a concept specifying the presence of an oriented edge at a particular location in the visual field has a low-order purview, whereas a concept specifying the extent of the entire visual field has a high-order purview. Concepts can also have low- and high invariance; for example, the concept for the letter A has high invariance because its purview specifies a high-order disjunction of states of the PSC elements (a specific arrangement of oriented edges in any of a large number of possible locations). Mechanisms specifying invariant concepts form a hierarchy going from low- to high-level areas of the cerebral cortex, as indicated by experimental data[52] and consistent with computational models for the recognition of objects[53], places, events[54] and spatial reference frames[55]. A concept can have low or high selectivity, depending on how strongly the state of its mechanism constrains its cause–effect repertoire. In the brain, the adaptive bias towards sparse firing makes it likely that the neurons would fire strongly when specifying a high invariance, high selectivity concept, such as the presence of the letter A (that is, a positive concept), and be silent when specifying its low selectivity counterpart, such as the absence of the letter A (that is, a negative concept) (FIG. 4).

In experimental settings, the content of experience is typically probed by asking the participant about high invariance, positive concepts, such as letters in the Sperling paradigm. However, we could undoubtedly report many more concepts than just the identity of a few letters. For example, we could report that there are many black symbols, that they are arranged in three rows and four columns, in a rectangular array, within a rectangular display, over a white homogeneous background that is spatially extended, being composed of a multitude of distinguishable locations, each with its specific neighbours, and so on. We can also report many negative concepts — for example, that the Sperling display did not include a face, a tree, an animal, a house, and so on — for the thousands of high invariance concepts we possess that happen to be negative for this particular image. Finally, we can report how all these concepts are bound together within the same experience in a complex pattern of relations — for example, we see the letter A as an invariant that is nevertheless located at a particular spatial location, that is composed of two oblique edges and a horizontal edge in between, that is capital, printed in black and located on the rightmost column in the upper row of the array, and so on. According to IIT, this dynamic binding of phenomenal attributes[56] occurs if, and only if, in cause–effect space the corresponding concept purviews are related, meaning that they refer to an overlapping set of PSC elements and jointly constrain their past or future states.

In short, the information that specifies an experience is much larger than the purported limited capacity of consciousness[57]. Although we are accustomed to summarizing what we see by referring to a few positive, high invariance concepts (for example, in FIG. 4 bottom panel, a participant may state: "I see the letters O, S and A"), we would not see what we see without the contribution of a large number of other concepts — low and high order, low and high invariance, positive and negative — and relations, which make the experience what it is (information) and thereby different from others (differentiation; FIG. 4). Consider what it would be like to look at the Sperling display not as a human, but as a machine implementing an efficient feedforward algorithm for letter recognition. The machine could certainly report three letters (in fact, all 12). However, such a machine could not see the scene and would understand virtually nothing because it has no other concept apart from the letters, not for the letter combination OSA, the array, the display, a face, an animal, and so on.

Indeed, if there were a face, an animal, or anything else in the middle of the display, it would do its best to categorize it as a letter.

Explanations

IIT provides a principled explanation for several seemingly disparate facts about the PSC. For example, IIT can explain why the cerebral cortex is important for consciousness, but the cerebellum is not. In general, the coexistence of functional specialization and integration in the cerebral cortex is ideally suited to integrating information (Supplementary information S3 (figure)). Specifically, the grid-like horizontal connectivity among neurons in topographically organized areas in the posterior cortex, augmented by converging–diverging vertical connectivity linking neurons along sensory hierarchies, should yield high values of Φ^{max}. By contrast, cerebellar micro-zones that process inputs and produce outputs that are feedforward and largely independent of each other cannot form a large complex; nor can they be incorporated into a cortical high Φ^{max} complex, even though each cerebellar micro-zone may be functionally connected with a portion of the cerebral cortex (Supplementary information S3 (figure))[1]. In principle, these differences in organization can explain why lesions of the cerebellum, which has four times more neurons than the cerebral cortex[58], do not seem to affect consciousness[7,8]. Furthermore, circuits providing inputs and outputs to a major complex may not contribute to consciousness directly. This seems to be true with neural activity in the peripheral sensory and motor pathways, as well as within circuits looping out and back into the cortex through the basal ganglia[59–61], despite their manifest ability to affect cortical activity and thereby to influence the content of experience indirectly (Supplementary information S3 (figure)).

IIT also accounts for the fading of consciousness during slow wave sleep when cortical neurons fire but, as a result of changes in neuromodulation, become bistable — that is, any input quickly triggers a stereotypical neuronal down-state, after which neurons enter an up-state and activity resumes stochastically[29]. Bistability implies a generalized loss of both selectivity (causal convergence or degeneracy) and effectiveness (causal divergence or indeterminism)[18] that results in a breakdown of information integration (Supplementary information S3 (figure)).

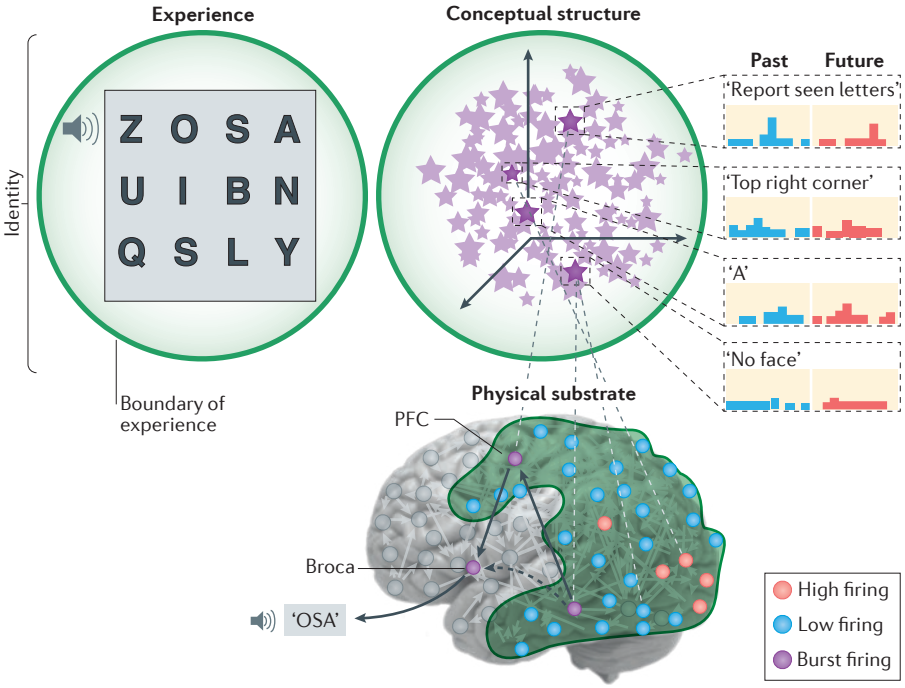


Figure 4 | **Phenomenal content and access content.** The content of an experience is much larger than what can be reported by a subject at any point in time. The left-hand panel illustrates the Sperling task[41], which involves the brief presentation of a three by four array of letters on a screen, and a particular row being cued by a tone. Out of the 12 letters shown on the display, participants correctly report only three or four letters — the letters cued by the tone — reflecting limited access. The top middle panel illustrates a highly simplified conceptual structure that corresponds to seeing the Sperling display, using the same conventions as outlined in FIG. 1. The myriad of positive and negative, first- and high-order, low- and high invariance concepts (represented by stars) that specify the content of this particular experience (seeing the Sperling display and having to report which letters were seen) make it what it is and different from countless other experiences (rich phenomenal content). The bottom panel schematically illustrates the physical substrate of consciousness (PSC) that might correspond to this particular conceptual structure (its boundary is represented by a green line). The PSC consists of neuronal groups that can be in a low firing state, a high firing state or a bursting state. Alone and in combination, these neuronal groups specify all the concepts that compose the conceptual structure. Stars that are linked to the PSC by grey dashed lines represent a small subset of these concepts. The PSC is synaptically connected to neurons in Broca's area by means of a limited capacity channel (dashed black arrow) that is dynamically gated by top-down connections (shown as solid black arrows) originating in the prefrontal cortex to carry out the instruction (that is, to report the observed letters 'OSA').

Findings from a study that used intracranial stimulation and recordings in patients with epilepsy are consistent with this account (Supplementary information S4 (box))[62]. During wakefulness, electrical stimulation of the cortex triggered a chain of deterministic phase-locked activations, whereas during slow wave sleep the same input induced a stereotyped slow wave that was associated with a cortical down-state (that is, a suppression of power ≥20 Hz). The cortical activity resumed to wakefulness-like levels after the down-state, but the phase-locking to the stimulus was lost, indicative of a break in the cause–effect chain (Supplementary information S4 (box)). Similar considerations would explain why information integration is impaired when consciousness fades despite the increased level of activity and synchronization that occurs early during generalized seizures[63].

IIT also provides a plausible account as to why conscious brains might have evolved. The world is immensely complex, at multiple spatial and temporal scales, and organisms with brains that can incorporate statistical regularities that reflect the causal structure of the environment into their own causal structure have an adaptive advantage for prediction and control[2]. The IIT framework, which emphasizes the information matching between intrinsic and extrinsic causal structures, has both similarities and differences with Bayesian approaches (for example, see REF. 64). According to IIT, given the constraints on energy and

space, organisms with brains of high Φ^{max} should have an adaptive advantage over less integrated competitors because they can fit more concepts (that is, functions) within a given number of neurons and connections. Simulated organisms (known as animats), whose 'brains' evolve by natural selection, show a monotonic relationship between integrated information and adaptation when placed in a maze[65]. Similarly, in the brain of animats that evolved to catch falling blocks in a simulated two-dimensional environment, both Φ^{max} and the number of concepts increased as a function of how well the animats performed on the task. Although in simpler environments animats with modular feedforward brains can catch blocks just as well, only animats with a high Φ^{max} evolve to adapt to more complex environments[66].

Predictions

At the most general level, IIT predicts that the PSC in the brain — that is, the major complex — must be a maximum of intrinsic cause–effect power, regardless of the particular set of neurons that constitute it (FIG. 3). IIT also predicts that the spatio-temporal grain of the physical elements specifying consciousness is that yielding the maximum Φ (FIG. 2). Testing these predictions experimentally is challenging but not impossible.

During the initial formulation of IIT, a systematic set of experiments was designed to test its specific prediction that consciousness requires both integration and differentiation[67]. An empirical measure, the perturbational complexity index (PCI), which can gauge the intrinsic cause–effect power of the cortex, has been introduced as a practical proxy for Φ^{max} (REF. 68). Calculating the PCI involves two steps: perturbing the cerebral cortex using transcranial magnetic stimulation to engage deterministic interactions among distributed groups of cortical neurons (integration) and measuring the incompressibility (algorithmic complexity) of the resulting responses (information). The PCI is high only if brain responses are both integrated and differentiated, corresponding to a distributed spatio-temporal pattern of causal interactions that is complex and hence not very compressible. So far, studies using PCI have confirmed the prediction of IIT that the loss and recovery of consciousness is associated with the breakdown and recovery of the capacity for information integration. This relationship holds true across different states of sleep[69] and anaesthesia (using different agents with various mechanisms of action)[70] and in patients with brain damage, at the level of single subjects[68]. Importantly, once PCI is validated in participants that can report on whether they were conscious or not, the index can be used to assess the capacity for information integration in patients who are unresponsive (such as those in a vegetative state) or cannot report (such as newborn infants and non-human species).

Another approach to estimating differentiation and integration in practice is to investigate the average properties of neural interactions based on a representative sample of neural states that span many regions of cause–effect space, such as those triggered by a movie sequence[23]. The data from a candidate set of neural elements (for example, functional MRI blood oxygen level-dependent values) can then be analysed using measures of differentiation and integration based on the postulates of IIT[23]. It is also possible to obtain an indication of information capacity from the dynamics of spontaneous activity[26,71,72]. Some studies in rats[73], monkeys[74] and humans[75] have confirmed that the differentiation of blood oxygen level-dependent activity patterns decreases when consciousness is lost. A similar approach can be used to evaluate information matching — how well the intrinsic cause–effect structures specified by the brain fit the causal structure of the environment[2,23].

Similar approaches could also be used to test the prediction that consciousness should split if a single major complex splits into two or more complexes, and that the split should happen precisely when two maxima of integrated information supplant a single maximum. For example, we could progressively reduce the efficacy of transmission in the callosal fibres by cooling or by the use of optogenetics. IIT predicts that there would be a moment at which, as a result of a minor change in the traffic of neural impulses across the callosum, a single consciousness would suddenly split into two. As discussed earlier, a split from a single major complex into two or more might also be observed in functional blindness (when a patient claims to be blind but may purposefully avoid obstacles) and other dissociative disorders, perhaps even in healthy participants under certain circumstances (such as during autopilot-like driving while having a conversation) (FIG. 3f).

Turning to the contents of consciousness, the fundamental identity of IIT implies that all qualitative features of experience correspond to features of the conceptual structure specified by the PSC. For example, the organization of experience into distinct modalities (such as sight, hearing and touch) and submodalities (such as colour, shape and motion within the modality of sight) should correspond to the presence, within a conceptual structure, of distinct sets of concepts with extensively overlapping purviews within each set, but much less across sets[2]. IIT further predicts that the binding[56] of phenomenal distinctions, such as seeing a blue book on the piano on the left, should correspond, in the conceptual structure, to an overlap in the purview of the respective concepts (a relation). Also, differences between experiences should correspond to distances among conceptual structures in cause–effect space and dissimilarities among phenomenal distinctions within an experience should correspond to distances between concepts. The refinement of experience that occurs through learning (for example, learning to discriminate the taste of different wines) should be reflected in a refinement of shapes in cause–effect space as a result of the addition and splitting of concepts.

IIT also predicts that the spatial structure that characterizes much of our daily experience should be reflected in features of conceptual structures that are specified by connections among neurons arranged in two-dimensional grids. For example, horizontal connections within topographically organized visual areas would be needed to experience visual space from the intrinsic perspective, rather than merely serving to mediate modulatory contextual effects. This also implies that local strengthening or weakening of such horizontal connections in topographic areas should lead to a local distortion of experienced visual space, even though the feedforward mapping of visual inputs from the world remains unchanged.

More generally, IIT predicts that changes in the efficacy of the connections among elements of the PSC should lead to changes in experience even when these changes are not accompanied by changes in activity. A counterintuitive consequence of this prediction is that a brain area could contribute to an experience even if it is inactive but not if its connections or neurons are inactivated. Thus topographic visual areas would create visual space even in the absence of spiking activity but not if the horizontal connections within those areas are inactivated. Similarly, if the connections of neurons in colour areas are intact, the neurons would contribute to experience even if they are silent, by

specifying negative colour concepts, such as when seeing a picture in black and white. However, if the connections are damaged, they would not specify any colour concepts, as with certain achromatopsic patients who do not even understand that the picture is missing colour[27] (FIG. 3c). Similarly, IIT predicts that the cerebral cortex as a whole may support experience even if it is almost silent, a state which may perhaps be reached through meditative practices designed to achieve 'naked awareness' without content[76]. This contrasts with the common assumption that neurons only contribute to consciousness if they are active and 'broadcast' the information they represent[77,78] (Supplementary information S5 (box)). States of naked awareness could be compared with states of unawareness that occur, for example, during deep sleep or anaesthesia, when the cause–effect repertoires of cortical neurons, regardless of the level of neuronal activity, are disrupted as a result of bistability[79].

Conclusions

In summary, IIT is a theory of consciousness that starts from the self-evident, essential properties (axioms) of experience and translates them into the necessary and sufficient conditions (postulates) for the PSC. The axioms are intrinsic existence (my experience exists from my own intrinsic perspective); composition (it has structure), information (it is specific), integration (it is unitary) and exclusion (it is definite). The corresponding postulates state that the physical substrate of an experience must have cause–effect power upon itself (intrinsic existence); its parts must have cause–effect power within the whole (composition); and the cause–effect power of the PSC must be specific (information), irreducible (integration) and maximally so (exclusion). The fundamental identity of IIT states that the quality or content of consciousness is identical to the form of the conceptual structure specified by the PSC, and the quantity or level of consciousness corresponds to its irreducibility (integrated information Φ).

The assessment of the identity between experiences and conceptual structures as proposed by IIT is clearly a demanding task, not only experimentally, but also mathematically and computationally. Evaluating maxima of intrinsic cause–effect power systematically requires going through many levels of organization, at multiple temporal scales, in many sets of brain regions, while performing an extraordinary number of perturbations and observations. Hopefully, heuristic approaches will be sufficient to make a strong case that the PSC is constituted of some particular neural elements, timescales and activity states. It will then be essential to test the prediction that any manipulation that affects the PSC at the spatio-temporal grain of maximum intrinsic cause–effect power should affect experience. Conversely, similar manipulations that do not affect the PSC, or that affect it at the wrong spatio-temporal grain, should leave experience unchanged. These and other predictions, especially those that are counterintuitive, will also help in assessing the validity of IIT in relation to other proposals about the neural basis of consciousness (Supplementary information S5 (box)).

Importantly, the more convincingly IIT can be validated under conditions in which it is relatively easy to assess how consciousness changes, the more it will help to make inferences about consciousness in hard examples, such as brain-damaged patients with residual areas of cortical activity, fetuses, infants, animals and machines. If it is validated, IIT may also prompt a reconsideration of how widespread consciousness is in nature and at what physical scale it may occur[9]. Intriguingly, IIT allows for certain simple systems such as grid-like architectures, similar to topographically organized areas in the human posterior cortex, to be highly conscious even when not engaging in any intelligent behaviour. Conversely, digital computers running complex programs based on a von Neumann architecture would not be conscious, even though they may perform highly intelligent functions and simulate human cognition. IIT offers a principled, empirically testable and clinically useful account of how three pounds of organized, excitable matter support the central fact of our existence — subjective experience. Time will tell whether this account is anywhere near the mark.

Giulio Tononi is at the Department of Psychiatry, University of Wisconsin, 6001 Research Park Boulevard, Madison, Wisconsin 53719, USA.

Melanie Boly is at the Department of Psychiatry, University of Wisconsin, 6001 Research Park Boulevard, Madison, Wisconsin 53719 USA; and at the Department of Neurology, University of Wisconsin, 1685 Highland Avenue, Madison, Wisconsin 53705, USA.

Marcello Massimini is at the Department of Biomedical and Clinical Sciences 'Luigi Sacco', University of Milan, Via G.B. Grassi 74, Milan 20157, Italy; and at the Instituto Di Ricovero e Cura a Carattere Scientifico, Fondazione Don Carlo Gnocchi, Via A. Capecelatro 66, Milan 20148, Italy.

Christof Koch is at the Allen Institute for Brain Science, 615 Westlake Ave N, Seattle, Washington 98109, USA.

Correspondence to G.T.
gtononi@wisc.edu

doi:10.1038/nrn.2016.44
Published online 26 May 2016

1. Oizumi, M., Albantakis, L. & Tononi, G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* **10**, e1003588 (2014).
2. Tononi, G. The integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* **150**, 56–90 (2012).
3. Tononi, G. Integrated information theory. *Scholarpedia* http://dx.doi.org/10.4249/scholarpedia.4164 (2015).
4. Posner, J. B., Saper, C. B., Schiff, N. D. & Plum, F. *Diagnosis of Stupor and Coma* (Oxford Univ. Press, 2007).
5. Koch, C., Massimini, M., Boly, M. & Tononi, G. The neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* **17**, 307–321 (2016).
6. Boly, M. *et al.* Consciousness in humans and non-human animals: recent advances and future directions. *Front. Psychol.* **4**, 625 (2013).
7. Lemon, R. N. & Edgley, S. A. Life without a cerebellum. *Brain* **133**, 652–654 (2010).
8. Yu, F., Jiang, Q. J., Sun, X. Y. & Zhang, R. W. A new case of complete primary cerebellar agenesis: clinical and imaging findings in a living patient. *Brain* **138**, e353 (2015).
9. Tononi, G. & Koch, C. Consciousness: here, there, and everywhere? *Phil. Trans. R. Soc. B* **370**, 20140167 (2015).
10. Chalmers, D. J. Facing up to the problem of consciousness. *J. Conscious. Studies* **2**, 200–219 (1995).
11. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **5**, 42 (2004).
12. Tononi, G. Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* **215**, 216–242 (2008).
13. Descartes, R. *Discourse on Method and Meditations on First Philosophy* (Hackett, 1998).
14. Pöppel, E. *Mindworks: Time and Conscious Experience* (Harcourt Brace Jovanovich, 1988).
15. Holcombe, A. O. Seeing slow and seeing fast: two limits on perception. *Trends Cogn. Sci.* **13**, 216–221 (2009).
16. Bachmann, T. *Microgenetic Approach to the Conscious Mind* (John Benjamins, 2000).
17. Kim, J. Multiple realization and the metaphysics of reduction. *Philos. Phenomenol. Res.* **52**, 1–26 (1992).
18. Hoel, E. P., Albantakis, L. & Tononi, G. Quantifying causal emergence shows that macro can beat micro. *Proc. Natl Acad. Sci. USA* **110**, 19790–19795 (2013).
19. Alivisatos, A.P. *et al.* The brain activity map project and the challenge of functional connectomics. *Neuron* **74**, 970–974 (2012).
20. Buzsáki, G. Neural syntax: cell assemblies, synapsembles, and readers. *Neuron* **68**, 362–385 (2010).
21. Li, C. Y., Poo, M. M. & Dan, Y. Burst spiking of a single cortical neuron modifies global brain state. *Science* **324**, 643–646 (2009).
22. London, M., Roth, A., Beeren, L., Häusser, M. & Latham, P. E. Sensitivity to perturbations *in vivo* implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
23. Boly, M. *et al.* Stimulus set meaningfulness and neurophysiological differentiation: a functional magnetic resonance imaging study. *PLoS ONE* **10**, e0125337 (2015).
24. Boly, M. *et al.* Brain connectivity in disorders of consciousness. *Brain Connect.* **2**, 1–10 (2012).
25. Seth, A. K., Barrett, A. B. & Barnett, L. Causal density and integrated information as measures of conscious level. *Philos. Trans. A Math. Phys. Eng. Sci.* **369**, 3748–3767 (2011).
26. Deco, G., Hagmann, P., Hudetz, A. G. & Tononi, G. Modeling resting-state functional networks when the cortex falls asleep: local and global changes. *Cereb. Cortex* **24**, 3180–3194 (2014).
27. von Arx, S. W., Muri, R. M., Heinemann, D., Hess, C. W. & Nyffeler, T. Anosognosia for cerebral achromatopsia — a longitudinal case study. *Neuropsychologia* **48**, 970–977 (2010).

28. Goldberg, I. I., Harel, M. & Malach, R. When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron* **50**, 329–339 (2006).
29. Steriade, M., Timofeev, I. & Grenier, F. Natural waking and sleep states: a view from inside neocortical neurons. *J. Neurophysiol.* **85**, 1969–1985 (2001).
30. Nir, Y. *et al.* Regional slow waves and spindles in human sleep. *Neuron* **70**, 153–169 (2011).
31. Siclari, F., LaRocque, J. J., Bernardi, G., Postle, B. R. & Tononi, G. The neural correlates of consciousness in sleep: a no-task, within-state paradigm. *BioRXiv* http://dx.doi.org/10.1101/012443 (2014).
32. Sperry, R. W. in *Neuroscience 3rd Study Program* (eds Schmitt, F. O. & Worden, F. G.) 5–19 (MIT Press, 1974).
33. Gazzaniga, M. S. Forty-five years of split-brain research and still going strong. *Nat. Rev. Neurosci.* **6**, 653–659 (2005).
34. Berlin, H. A. The neural basis of the dynamic unconscious. *Neuropsychoanalysis* **13**, 1–68 (2011).
35. Mudrik, L., Breska, A., Lamy, D. & Deouell, L. Y. Integration without awareness: expanding the limits of unconscious processing. *Psychol. Sci.* **22**, 764–770 (2011).
36. Mudrik, L., Faivre, N. & Koch, C. Information integration without awareness. *Trends Cogn. Sci.* **18**, 488–496 (2014).
37. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
38. Harris, K. D. & Shepherd, G. M. The neocortical circuit: themes and variations. *Nat. Neurosci.* **18**, 170–181 (2015).
39. Miller, G. A. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**, 81–97 (1956).
40. Norretranders, T. *The User Illusion: Cutting Consciousness Down to Size* (Viking Penguin, 1991).
41. Sperling, G. The information available in brief visual presentations. *Psychol. Monogr.* **74**, 1–29 (1960).
42. Cohen, M. A. & Dennett, D. C. Consciousness cannot be separated from function. *Trends Cogn. Sci.* **15**, 358–364 (2011).
43. Cohen, M. A. & Dennett, D. C. Response to Fahrenfort and Lamme: defining reportability, accessibility and sufficiency in conscious awareness. *Trends Cogn. Sci.* **16**, 139–140 (2012).
44. O'Regan, J. K., Rensink, R. A. & Clark, J. J. Change-blindness as a result of 'mudsplashes'. *Nature* **398**, 34–34 (1999).
45. Dehaene, S. *Consciousness and the Brain: Deciphering How the Brain Codes our Thoughts* (Penguin, 2014).
46. Kouider, S., de Gardelle, V., Sackur, J. & Dupoux, E. How rich is consciousness? The partial awareness hypothesis. *Trends Cogn. Sci.* **14**, 301–307 (2010).
47. Block, N. On a confusion about a function of consciousness. *Behav. Brain Sci.* **18**, 227–287 (1995).
48. Block, N. Perceptual consciousness overflows cognitive access. *Trends Cogn. Sci.* **15**, 567–575 (2011).
49. Lamme, V. A. How neuroscience will change our view on consciousness. *Cogn. Neurosci.* **1**, 204–220 (2010).
50. Bronfman, Z. Z., Brezis, N., Jacobson, H. & Usher, M. We see more than we can report: "cost free" color phenomenality outside focal attention. *Psychol. Sci.* **25**, 1394–1403 (2014).
51. Wolfe, J. in *Fleeting Memories* (ed. Coltheart, V.) 71– 94 (MIT Press, 2000).
52. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
53. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
54. Franzius, M., Sprekeler, H. & Wiskott, L. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* **3**, e166 (2007).
55. Spratling, M. W. Learning posture invariant spatial representations through temporal correlations. *IEEE Trans. Autonom. Ment. Dev.* **1**, 253–263 (2009).
56. Treisman, A. The binding problem. *Curr. Opin. Neurobiol.* **6**, 171–178 (1996).
57. Baddeley, A. D. *Working Memory* (Clarendon Press, 1986).
58. Herculano-Houzel, S. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proc. Natl Acad. Sci. USA* **109** (Suppl. 1), 10661–10668 (2012).
59. Jain, S. K. *et al.* Bilateral large traumatic basal ganglia haemorrhage in a conscious adult: a rare case report. *Brain Inj.* **27**, 500–503 (2013).
60. Straussberg, R. *et al.* Familial infantile bilateral striatal necrosis: clinical features and response to biotin treatment. *Neurology* **59**, 983–989 (2002).
61. Caparros-Lefebvre, D., Destee, A. & Petit, H. Late onset familial dystonia: could mitochondrial deficits induce a diffuse lesioning process of the whole basal ganglia system? *J. Neurol. Neurosurg. Psychiatry* **63**, 196–203 (1997).
62. Pigorini, A. *et al.* Bistability breaks-off deterministic responses to intracortical stimulation during non-REM sleep. *Neuroimage* **112**, 105–113 (2015).
63. Blumenfeld, H. Impaired consciousness in epilepsy. *Lancet Neurol.* **11**, 814–826 (2012).
64. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
65. Edlund, J. A. *et al.* Integrated information increases with fitness in the evolution of animats. *PLoS Comput. Biol.* **7**, e1002236 (2011).
66. Albantakis, L., Hintze, A., Koch, C., Adami, C. & Tononi, G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Comput. Biol.* **10**, e1003966 (2014).
67. Massimini, M. *et al.* Breakdown of cortical effective connectivity during sleep. *Science* **309**, 2228–2232 (2005).
68. Casali, A. G. *et al.* A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl Med.* **5**, 198ra105 (2013).
69. Massimini, M. *et al.* Cortical reactivity and effective connectivity during REM sleep in humans. *Cogn. Neurosci.* **1**, 176–183 (2010).
70. Sarasso, S. *et al.* Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Curr. Biol.* **25**, 3099–3105 (2015).
71. Barrett, A. B. & Seth, A. K. Practical measures of integrated information for time-series data. *PLoS Comput. Biol.* **7**, e1001052 (2011).
72. Oizumi, M., Amari, S., Yanagawa, T., Fujii, N. & Tsuchiya, N. Measuring integrated information from the decoding perspective. *PLoS Comput Biol* **12**, e1004654 (2015).
73. Hudetz, A. G., Liu, X. & Pillay, S. Dynamic repertoire of intrinsic brain states is reduced in propofol-induced unconsciousness. *Brain Connect.* **5**, 10–22 (2015).
74. Barttfeld, P. *et al.* Signature of consciousness in the dynamics of resting-state brain activity. *Proc. Natl Acad. Sci. USA* **112**, 887–892 (2015).
75. Tagliazucchi, E. *et al.* Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J. R. Soc. Interface* **13**, 20151027 (2016).
76. Sullivan, P. R. Contentless consciousness and information-processing theories of mind. *Philos. Psychiatry Psychol.* **2**, 51–59 (1995).
77. Baars, B. A. *Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1988).
78. Dehaene, S. & Changeux, J.-P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227 (2011).
79. Steriade, M. The corticothalamic system in sleep. *Front. Biosci.* **8**, d878-99 (2003).
80. Searle, J. Can information theory explain consciousness? *New York Review of Books* (10 Jan 2013).

Acknowledgements

The authors thank L. Albantakis, C. Cirelli, L. Ghilardi, W. Marshall, W. Mayner, A. Mensen, M. Oizumi, U. Olcese, B. Postle, S. Sasai and other colleagues for their various contributions to the work presented here. This work was supported by the Templeton World Charity Foundation, the McDonnell Foundation and the Distinguished Chair in Consciousness Science (University of Wisconsin) (to G.T.), and by the James S. McDonnell Scholar Award 2013 (to M.M.).

Competing interests statement

The authors declare no competing interests.

FURTHER INFORMATION

Integrated Information Theory: http://www.integratedinformationtheory.org

SUPPLEMENTARY INFORMATION

See online article: S1 (figure) | S2 (box) | S3 (figure) | S4 (box) | S5 (box)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF