Trends in
Cognitive Sciences

CellPress
OPEN ACCESS

Opinion

The strength of weak integrated information theory

Pedro A.M. Mediano,[1,2,*] Fernando E. Rosas,[3,4,5] Daniel Bor,[1,2] Anil K. Seth,[6,7] and Adam B. Barrett[6,8,*]

The integrated information theory of consciousness (IIT) is divisive: while some believe it provides an unprecedentedly powerful approach to address the 'hard problem', others dismiss it on grounds that it is untestable. We argue that the appeal and applicability of IIT can be greatly widened if we distinguish two flavours of the theory: strong IIT, which identifies consciousness with specific properties associated with maxima of integrated information; and weak IIT, which tests pragmatic hypotheses relating aspects of consciousness to broader measures of information dynamics. We review challenges for strong IIT, explain how existing empirical findings are well explained by weak IIT without needing to commit to the entirety of strong IIT, and discuss the outlook for both flavours of IIT.

Highlights

The integrated information theory of consciousness (IIT) is unprecedentedly ambitious in that it proposes a universal mathematical formula, derived from fundamental properties of conscious experience, to describe the quality and quantity of consciousness for any physical system that possesses it.

IIT proponents believe it may solve the 'hard problem' of consciousness of why and how physical processes can be accompanied by subjective experience.

However, in the current formulation, IIT formulae are not always well-defined and current empirical evidence does not support the level of specificity present in the theory.

At the same time, available empirical evidence does support a weaker, less prescriptive version of the theory.

We argue that distinguishing a 'weak' from a 'strong' flavour of IIT can provide a useful theoretical umbrella for ongoing empirical work, widening the overall appeal and applicability of the theory.

Divide and conquer

IIT (see Glossary) has gained considerable prominence among theories of consciousness, in large part because of its ambitious claim to specify the necessary and sufficient basis for any physical substrate of consciousness [1]. The theory proposes a mathematical formula, derived by distilling the fundamentals of phenomenology into a small set of axioms, which is posited to describe the quantity and quality of the consciousness for any physical system that possesses it [1,2]. Furthermore, practically applicable IIT-inspired measures of the 'complexity' of neural dynamics [3–5] behave in concordance with the predictions of IIT and have found important clinical application in the assessment of conscious level in brain injury patients suffering disorders of consciousness [6,7]. However, the fundamental formula posited by IIT is intractable, except in certain 'toy' systems, and is ill-defined in some cases (including in the key application case of the human brain) [8,9]. Thus, existing IIT-inspired measures do not provide specific tests of IIT (e.g., tests that distinguish IIT from other possible similar theories); instead, they demonstrate correlations between certain aspects of macroscopic neural activity and the level of consciousness, which have also been recognised as potentially relevant by other theoretical frameworks [10,11]. The combination of IIT's ambitious claims with these difficulties for testing its most specific, distinctive claims have generated considerable confusion and polarisation [12,13], as well as criticism [14–16]. With large international projects now underway attempting to pit IIT against competing theories of consciousness [17], it is crucial to clarify the landscape surrounding IIT so that empirical research can be better matched to theoretical claims.

Here, we propose that the appeal and applicability of IIT can be widened by distinguishing between 'strong' and 'weak' flavours of the theory. Strong/weak distinctions have a long history in science, with two prominent examples being artificial intelligence [18] and emergence [19–21]. Broadly, a 'strong' perspective tends to have an ontological flavour, prescribing how things are; whereas a 'weak' perspective aims to describe a phenomenon, by explaining and simulating its properties. These distinctions are not only conceptually useful but also scientifically productive, as

[1]Department of Psychology, University of Cambridge, Cambridge, UK
[2]Department of Psychology, Queen Mary University of London, London, UK
[3]Centre for Psychedelic Research, Imperial College London, London, UK
[4]Data Science Institute, Imperial College London, London, UK
[5]Centre for Complexity Science, Imperial College London, London, UK
[6]Sackler Centre for Consciousness Science, Department of Informatics, University of Sussex, Brighton, UK
[7]CIFAR Program on Brain, Mind, and Consciousness, Toronto, Canada
[8]The Data Intensive Science Centre, Department of Informatics, University of Sussex, Brighton, UK

*Correspondence:
pam83@cam.ac.uk (P.A.M. Mediano)
and abb22@sussex.ac.uk (A.B. Barrett).

they enable scientists with a broader range of philosophies, objectives, and approaches to engage with the theories in question.

In the context of IIT, these distinctions play out as follows (see Table 1 for an itemised view). In **strong IIT**, states of consciousness are identified with maxima of integrated information in any physical system. By contrast, **weak IIT** will test pragmatic hypotheses based on **explanatory correlates** [22] between the dynamics of information **integration** and certain aspects of consciousness. Strong IIT considers consciousness to be a fundamental universal physical property, such as charge or mass [23], and assumes that a universally applicable formula for describing consciousness can in theory be obtained, with phenomenology and theoretical physics as the drivers in its construction. Weak IIT is agnostic on whether this is the case and hence can accommodate a broader range of philosophical perspectives. Moreover, hypotheses generated by weak IIT circumvent the tractability issues of the strong approach and can be directly formulated for empirically observable neurophysiological variables. In practice, strong IIT emphasises theoretical developments to inform new measures, while weak IIT focuses on applications of measures to guide theory. Together, they can foster complementary and mutually enriching research programmes.

After a brief overview of strong IIT and its theoretical challenges, this article outlines the principles behind weak IIT and sets out the advantages it offers over strong IIT, especially in terms of interpretation of empirical data. We conclude by describing some possible paths ahead for weak IIT, for it to best contribute to the development of consciousness science.

Brief overview of strong IIT

Strong IIT [2] attempts to derive a universal formula for consciousness based on five fundamental properties of phenomenology, referred to as *axioms* [2,24]. (See Box 1 for a brief history of IIT.) The first property is *intrinsicality*, which says that experience is subjective, existing from the intrinsic perspective of the subject of experience. The second property, *composition*, states that experience is structured, being composed of several phenomenal distinctions that exist within it; for example, within a single experience one may distinguish a piano, a blue colour, a book, countless spatial locations, sounds, various emotions, and so on. Third, *information*, reflects that conscious experiences are informative, in the sense that each experience is specific and in some sense rules out other potential experiences that were *a priori* possible. *Integration* states that experience is unified, in that it cannot be subdivided into parts that are experienced separately. Finally, *exclusion* says that experience is definite, in that there do not exist simultaneous sets of experiences generated by overlapping physical systems. In addition there is *existence*, a ‘zeroth axiom’ that states that all substrates of consciousness must exist in physical terms. An innovative aspect of strong IIT is that it addresses the **hard problem** ‘backwards’, by proceeding from phenomenological axioms to mechanisms, as opposed to trying to go from mechanisms to consciousness.

Table 1. Key differences between weak and strong IIT

Strong IIT	Weak IIT
Addresses the ‘hard problem’ of consciousness [67]	Addresses the ‘real problem’ of consciousness [68]
Claims an identity between consciousness and maximally integrated cause–effect structures	Uses integrated information measures as explanatory correlates for properties of consciousness
Applies to all physical systems	Applies (so far) to neural systems only
Focuses on theoretically fundamental, rather than practical, measures	Focuses on practical measures for real brain data

Box 1. History of IIT

IIT grew out of the intuition that dynamical complexity, understood as coexistence of differentiation (the system having elements that are functionally and dynamically distinct) and integration (the system behaving coherently as a whole), ought to be a key feature of the neural activity associated with consciousness, since these properties are also general properties of (arguably) all conscious experiences [70,71]. This idea was first operationalised by the mutual information-based measure of *neural complexity* [72] and then by the first Φ measure in IIT 1.0, which was based on the number of possible states of the system and the statistical interdependencies between system components [73,74].

The second version (IIT 2.0) introduced a new Φ measure based on the information generated by the system as it transitions from one state to the next [23] – for the first time identifying consciousness with properties of dynamical transitions. At the same time, it emphasised the role of interelement causal connections as determinants of consciousness. Conscious contents, and the quality of consciousness, arise collectively from the informational relationships between the states of all subsets of the system [75], in principle describable by exploring Φ on all system subsets.

The most recent version (IIT 3.0) identifies three additional properties of phenomenology (*existence*, *composition*, and *exclusion*), adding them to the properties of differentiation (which was reframed as *information*) and integration, to extend the mathematical formalism of integrated information and formulate a new measure, Φ^{Max} [1,2]. While Φ^{Max} still provides a measure of the overall level of consciousness, the emphasis in IIT 3.0 is more on establishing a theoretical mapping between the cause–effect structure of a physical system and the structure of any conscious experience associated with it. Specifically, a system has a 'conceptual structure' that can be derived from the cause–effect relations between its elements and conscious experiences are identical to these conceptual structures. At the time of writing, IIT 4 is a work-in-progress [24] and contains *intrinsicality* as an additional 'axiom' (see main text).

As stated in its main articles [2,25], from these axioms, strong IIT posits *postulates* about the nature of the physical substrate of consciousness (PSC). From *existence*, *intrinsicality*, and *information*, a formula is constructed based on the probability of occurrence of each past and future state of the system given the current state, assuming that all states were equally likely *a priori* (i.e., in technical language, a 'maximum entropy' prior). Applying *composition*, the experience will depend on information provided by each system subset about the past and future of all subsets. And, applying *integration*, the experience depends on the extent to which whole subsets carry more information than nonoverlapping collections of their parts. From *exclusion* comes maximisations of integrated information: (i) over all discrete grainings of the system, in space, time, and the set of possible states of the system components; and (ii) over all system subsets. This process culminates in a measure of overall conscious level, $\boldsymbol{\Phi}^{\mathbf{Max}}$ (i.e., the 'amount' of consciousness).

In addition to specifying a formula for conscious level, strong IIT also studies the contents of consciousness in terms of the cause–effect structure of the physical substrate that gives rise to a maximum of integrated information (Box 1). Thus, strong IIT is a relatively comprehensive theory of consciousness, accounting for both the presence, degree, and character of conscious experiences. Importantly, though, all aspects of strong IIT rest on identifying maxima of integrated information in ways specified by the axioms and postulates.

The full formula for computing Φ^{Max} involves a considerable amount of mathematical detail [2,24], yet, as it currently stands, it has two major problems: it is not universally well-defined [8,9], and there is logical inconsistency in the postulates [26]. Two reasons (amongst others) that it is ill-defined are: (i) the so-called maximum entropy prior it relies on does not exist for systems with long 'memory' (i.e., non-Markovian systems, such as the brain [27], in which the transition between states depends on the full past history of the system); and (ii) the maximisations involved can result in ties, which leaves the remainder of the calculation ill-defined and leads to the so-called 'underdetermined qualia problem' (while this situation may be unlikely in practice, it is still a problem for a theory aiming to describe consciousness at a fundamental level) [9,28]. The logical inconsistency is that *intrinsicality* requires Φ^{Max} to be an intrinsic property, while *exclusion* is extrinsic, since it requires a maximisation of Φ^{Max} that involves comparisons with other systems

Glossary

Differentiation: property of a system whereby its elements are functionally and dynamically distinct.

Empirical Φ measures: measures of integrated information designed to be applicable to time series data. In general, they are based on statistics derived from what the system actually does, as opposed to what the system could possibly do.

Explanatory correlate: a neural process that both correlates with, and also accounts for, functional and phenomenological properties of consciousness. An explanatory correlate of some particular conscious process helps us understand why that conscious process has the functional and phenomenological character that it has.

Hard problem: the problem of explaining the relationship between physical phenomena, such as brain processes, and the conscious experiences (qualia) they generate. This is defined in contrast to the 'easy problems' of understanding, for example, how the brain categorises and discriminates environmental stimuli, or focuses attention.

Information atoms: constituent components of mutual information. Together, they provide an exhaustive multidimensional description of the information dynamics within a system.

Integrated information theory (IIT): theory of consciousness that posits a relationship of subjective experience with complex interactions between elements in a system (like the human brain).

Integration: property of a system whereby it behaves coherently as a whole.

Partial information decomposition: information-theoretic tool used to decompose mutual information into distinct information atoms, usually labelled as synergistic, redundant, and unique.

$\boldsymbol{\Phi}^{\mathbf{Max}}$: measure of integrated information in strong IIT, derived from its postulates.

Strong IIT: flavour of IIT as presented in its main papers, that identifies consciousness with.the cause–effect structure of a physical substrate that specifies a maximum of integrated information.

Weak IIT: flavour of IIT introduced here, that searches for robust explanatory correlations between aspects of consciousness and aspects of information dynamics.

[26]. Furthermore, due to the optimisations over coarse-grainings and system subsets, the complicated procedure to calculate Φ^{Max} becomes intractable in systems with any reasonable degree of mechanistic complexity [28].

Despite these issues, the core of strong IIT rests on a simple and elegant idea: that consciousness is identical to properties of intrinsic integrated information in a system. It remains possible that a mathematical formulation that addresses the aforementioned problems, based on similar principles to those currently set out, could in the future be plausible [29,30]. Work is ongoing on characterising probability distribution spaces in an intrinsic way [24,31]. Meanwhile, the debate continues about the veracity, consistency, and universality of the fundamental axioms [26,32].

Empirical tests of strong IIT are likely to remain challenging, precisely because of its high level of ambition to identify the PSC precisely and universally. While certain aspects of the nature of the PSC posited by strong IIT can be empirically tested to some extent (e.g., by considering the difference between inactive versus inactivated neurons to discern if, as strong IIT suggests [33], it is the brain's causal structure, beyond mere activity, that is responsible for consciousness), there is no existing experimental finding that favours the whole of strong IIT over mere components of it, or indeed over a weaker set of theoretical assumptions. Moreover, existing findings can be alternatively explained by a distinct theoretical framework, without committing to all the dramatic claims of strong IIT, namely weak IIT, to which we now turn.

Weak IIT

The goal of weak IIT, as outlined here, is to search for empirically measurable and powerful explanatory correlates of various aspects of consciousness. Weak IIT shares many motivations with strong IIT, but has a focus on practical measures for real brain data. In particular, weak IIT preserves the idea, central to IIT since its inception, that neural substrates of consciousness must reflect two key phenomenological observations: (i) each conscious moment is highly informative (it is one of a vast repertoire of possible experiences); and (ii) each conscious experience is integrated (it is experienced as a coherent whole) (Box 2). However, and crucially, weak IIT no longer claims an identity relationship; on weak IIT, integrated information is an explanatory correlate of consciousness, but it may not be a strictly necessary or sufficient condition for it.

Box 2. Integrated experience, integrated dynamics

The core theoretical argument behind weak IIT is that an integrated experience in the phenomenological sense should be generated by integrated brain activity in the statistical sense.

The argument rests on two assumptions. The first is that, phenomenologically, conscious experience is *integrated* (or unified), as all elements within it form a single cohesive whole and changing any one of them would change the experience altogether [76]. The second is that multiple aspects of any given experience (from shapes and colours to sounds and evoked memories) are encoded by different regions in the brain. Therefore, if these brain regions are to give rise to conscious experience, they should do so through statistical interactions spanning multiple brain regions [77].

Similarly, weak IIT assumes that a rich conscious life is generated, in part, by the large number of different experiences available. In other words, any particular subjective experience is *informative*, by virtue of ruling out many other possible alternative experiences. Dynamically, this translates to the statement that the brain must have access to and spontaneously visit a large and diverse repertoire of states mapping onto those possible experiences (as opposed to strong IIT, which only requires states to be accessible in principle, not necessarily visited in practice). This property is related to the pre-IIT concept of differentiation (cf. Box 1): a system cannot visit many states if all of its parts are fully correlated, therefore, some degree of differentiation is needed for the system to be informative.

Taken together, these two (brief) arguments suggest a link between consciousness and properties of neural dynamics: specifically, that the joint dynamics of brain regions must be highly diverse yet statistically interdependent.

Accordingly, weak IIT does not commit to all the axioms of strong IIT and does not make claims of generalisation beyond the brain.

Separating weak IIT from strong IIT helps interpret a range of existing empirical work that has previously been lumped together under a single IIT banner. Broadly, any practical measure of **differentiation**, information, and/or integration has the potential to contribute to (or test) weak IIT. In fact, much work has already been carried out on what could be considered weak IIT, via the application of such measures to various datasets [4,34–36]. Previously, many of these studies have been described as providing support for strong IIT; however, weak IIT offers a more parsimonious interpretation, since not all the axioms and postulates of strong IIT are needed to account for the results. Moreover, given the concerns raised over Φ^{Max} [8,9,26], it is problematic to consider Φ-like measures designed for experimental application as actually testing an approximation to strong IIT, beyond testing weak IIT.

Weak IIT allows researchers to work on developing and experimentally testing the core intuitions of IIT without committing to the more contentious claims central to strong IIT (especially its identity claim) and/or having to address the open mathematical problems with Φ^{Max}. Importantly, there are benefits here for strong IIT too: distinguishing the two flavours allows strong IIT to focus on honing its mathematical and theoretical basis, without immediately concerning itself with direct empirical testability. Next, we describe specific examples of how weak IIT readily accommodates existing empirical results and outline ideas for its future development.

Integrated information and perturbational complexity

A much drawn-upon source of empirical evidence adduced in support of IIT comes from a series of experiments that examine the electroencephalographic (EEG) response to transcranial magnetic stimulation (TMS) [3,4,37]. In a landmark paper [4], Casali and colleagues introduced the Perturbational Complexity Index (PCI), an IIT-inspired measure that quantifies the signal diversity of the EEG response to TMS, and showed that it is a reliable index of conscious level across a wide range of states of consciousness (non-rapid eye movement sleep, general anaesthesia induced by a variety of compounds, various major disorders of consciousness). Overall, these studies found that when subjects are unconscious, the brain's response to TMS is stereotypical across electrodes and/or remains local to the site of stimulation (at standard stimulus intensities), whereas the response of conscious subjects is more diverse across electrodes and spreads across larger regions of cortex. In this way, the probed neural dynamics when the participant is conscious appear both to be more diverse and to play out across a wider network, which implies greater differentiation and greater integration, suggesting a natural connection between PCI and IIT and representing an 'implicitly weak' approach to IIT, as highlighted in recent work [5,38].

While PCI has been presented as a proxy for Φ^{Max}, and its success has been described as support for IIT [1,25], PCI's general behaviour does not comply with the strict requirements of strong IIT for a measure of integrated information. PCI simply counts the number of distinct patterns in a binary representation of the response and normalises this number based on what would be expected for random data with the same level of overall activity. Thus, high PCI could in theory be obtained with a response that does not spread far (i.e., without integrated dynamics), as long as the response that does occur has a relatively high signal diversity (although this is unlikely when applied to real brains). Further, PCI does not explicitly incorporate all the postulates of strong IIT: it does not incorporate (i) exclusion, since it does not consider comparisons between system subsets or across system grainings; (ii) composition, since it does not consider interactions between system components; or (iii) existence of all elements, since only one of them (the TMS site) is causally perturbed (the existence postulate requires all elements to be perturbed to assess their cause–effect power). Additionally, despite some proof-of-principle examples in logic-gate systems relating signal diversity and Φ^{Max} [39], there exists no mathematical argument linking Φ^{Max} and PCI at the whole-brain level. Finally, PCI is computed with respect to a specific locus and intensity of brain perturbation via TMS, which is fundamentally different from the theoretical maximum-entropy perturbation from which Φ^{Max} is calculated (which accounts for the evolution from all theoretically possible system states and through all possible state transitions).

The fact that PCI and Φ^{Max} are actually very different does not undermine the fact that PCI does capture aspects of integration and differentiation, which correlate extremely well with conscious level. Therefore, although PCI does not constitute specific evidence for strong IIT, it does support the less prescriptive principles behind weak IIT.

Experimental results on empirical Φ measures

Another branch of work on weak IIT is concerned with so-called **empirical Φ measures** [40] (i.e., measures of integrated information that are applicable to sets of variables for which time series data are available). There are now several such measures in addition to the original empirical measure Φ^{WMS} [40], including Φ^* [41], $\tilde{\Phi}$ [42], and Φ_R [43], among others [44–47], each of which operationalises in a different way the extent to which the whole set of variables contains more information than the sum of its parts. A crucial difference between these measures and the strong IIT Φ^{Max}, which makes them applicable to experimental data, is that they measure information based on the empirically observed distribution of system states, as opposed to a hypothetical maximum-entropy prior in which all possible states are equally likely.

Direct experimental evidence for correlations between empirical Φ measures and level of consciousness is still scarce, a fact often overlooked in many critiques of IIT. There is evidence for some Φ measures decreasing during loss of consciousness, (e.g., Φ^* in local field potentials [48], $\tilde{\Phi}$ in EEG [49], and Φ_R in fMRI [11]). However, there are also some conflicting findings, for example, Lee *et al.* [50] report decreased Φ^{WMS} under anaesthesia in the gamma band, while Kim *et al.* [51] report lower Φ^* in the alpha band but higher Φ^* in all other frequency bands. To make sense of these mixed results, computational work has shown that even in simple systems these measures can behave very differently [45], so it is not surprising they differ when applied to real data too.

A weak approach to IIT is well positioned to tackle the challenges raised here. Through combined experimental and theoretical work, it will be possible to understand what exactly is being captured by any specific empirical measure of integrated information, as well as the effects of idiosyncrasies of neural data. Altogether, by acknowledging that there may be no universal Φ measure, weak IIT can make progress by examining how different empirical measures relate to each other, for example, by embedding them in a larger family of measures (in which the known measures are particular special cases) [44] and by relating them to different aspects of consciousness.

Towards a multidimensional characterisation of integrated information

PCI and existing empirical Φ measures are scalar (i.e., one-dimensional) quantities. While this makes for simple experimental applications, it also means that by themselves they are unable to capture the diverse phenomenological properties of consciousness. Conscious experience, even conscious level, is not a monolithic entity, but an intricate amalgamation of many processes [52]. The view of consciousness as an aggregate phenomenon has evolved, with ever richer multidimensional descriptions gaining prominence [53–55]. Beyond this phenomenological argument, other mathematical arguments suggest that there are multiple ways in which information can be integrated in a system and thus that it is possible that no canonical measure of integrated information may exist, even in principle [43], emphasising the need to embrace multidimensional frameworks. Strong IIT, as noted, has the resources, at least in principle, to account for the character (content) of consciousness, as well as for level (Box 1). What are the prospects for weak IIT?

A path for multidimensional weak IIT

A multidimensional approach to weak IIT would minimally involve separate measures of the extent to which neural dynamics are differentiated and the extent to which they are integrated. Many studies have employed some simplified form of this multidimensional approach, taking, for example, some measure of entropy as a proxy for differentiation and some measure of correlation (or mutual information) as a proxy for integration [35,36,56–59]. Such an approach has already enabled an exploration of scenarios that are more nuanced than complete loss of consciousness, such as the psychedelic state, which exhibits both an increase in signal diversity [58,59] and breakdown in functional connectivity [57]; a state of increased differentiation but decreased integration that does not neatly fit into a one-dimensional view of conscious level.

A mature weak IIT would likely also involve further dimensions, including 'weak' equivalents of other axioms in strong IIT. For example, one might quantify to what extent additional properties of phenomenology are reflected in neural dynamics, such as composition, or diversity of information, as distinct from overall quantity of information (*à la* IIT 3.0). Furthermore, there are many modes of information transfer within a complex system (Box 3) and each of these modes might usefully describe a distinct correlate of some aspect of consciousness [43]. Speculatively, different balances of these modes might correspond with different 'ways of experiencing' [53–55] (e.g., by characterising the difference between ordinary consciousness and altered consciousness during a psychedelic experience [60]). The recently proposed integrated information decomposition (ΦID), described next, provides a promising framework for exploring this idea.

Integrated information decomposition (ΦID)

One recent development aligned with weak IIT is the examination of empirical integrated information measures in terms of different modes of information transfer. The framework for this has been called ΦID [43], which is a generalisation of the theory of **partial information decomposition** (PID) [61] to dynamical systems. The core insight of PID is that components of a system can carry information in different modes (referred to as **information atoms**), categorised as redundant, unique, synergistic, or some combination of these (Box 3). Information is said to be redundant if it is contained in more than one component, unique if it is contained in precisely one component, and synergistic if it is present in a group of components, but not in any of them individually. As an example, consider our two eyes as sources of information about the world: in this case, information about colour would be redundant, as it can be obtained from either eye; while information about depth would be synergistic (when obtained from binocular disparity), as it can only be obtained from having both eyes open simultaneously. Building on these ideas, ΦID generalises PID in order to make it applicable to multivariate time series data (such as EEG or fMRI from multiple brain regions).

ΦID offers one fertile opportunity for developing a multidimensional weak IIT. By design, ΦID quantifies multiple types of information dynamics, thus enabling a suite of multidimensional measures that includes (but is not limited to) more conventional Φ measures [43]. On the theoretical side, ΦID provides a common and encompassing space of information dynamics, which allows researchers to formulate specific testable hypotheses relating dimensions of this space with dimensions of consciousness (e.g., linking synergistic storage with selfhood [60]). Furthermore, on the practical side, ΦID has been used to refine previous Φ measures and enhance their power to discriminate between states of consciousness [11].

Box 3. Information dynamics in complex systems

Here we consider the ways in which information can be propagated forward in time from the state of a system at one time step to the state of a system at a future time step. For simplicity, we consider just two components of a system, which we call 1 and 2. We can represent their redundant information as Red, synergistic information as Syn, and unique information as Un^1 and Un^2, respectively, for 1 and 2. This decomposition into four kinds of information then yields 4 × 4 = 16 different ways in which information can be propagated into the future: subsets of all four kinds of information can be propagated to all four kinds of information. These 16 modes can be gathered into six groups of qualitatively distinct phenomena [43]:

- Storage: information that remains in the same set of components.
- Copy: information that becomes duplicated.
- Transfer: information that moves between components.
- Erasure: information that was duplicated and is removed from one component.
- Downward causation: collective properties that influence individual component futures.
- Upward causation: collective properties that are influenced by individual components.

With respect to this taxonomy, it becomes clear that there are many modes of information integration. The whole could be considered to be propagating more information than the sum of the parts due to transfer modes like $\text{Un}^1 \rightarrow \text{Un}^2$, downward causation modes like $\text{Syn} \rightarrow \text{Un}^1$, synergistic storage modes like Syn → Syn, or some combination of the above (Figure I). Mediano *et al.* [43] show explicitly how different empirical Φ measures weight these different modes of information integration differently and that these measures are not simply different approximations of a unique concept of integration, but instead capture fundamentally different aspects of a system's information dynamics. Correspondingly, we hypothesise these modes may reflect distinctive aspects of the brain dynamics underlying different aspects of consciousness when computed on experimental neuroimaging data [60].

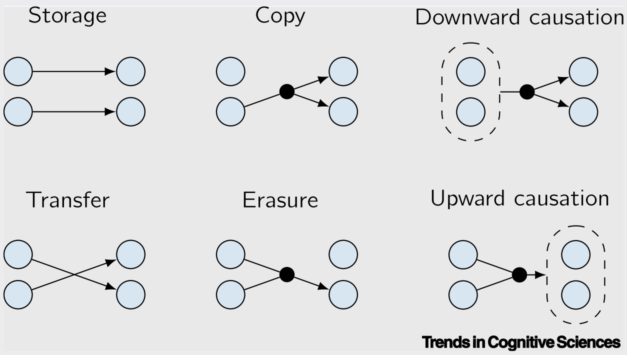

Figure I. Different modes of information dynamics illustrated in a bipartite system.

Outstanding questions

How should strong IIT address its open challenges? Should it develop a new mathematical framework (based, e.g., on category theory), or aim for a closer correspondence with fundamental physics?

What are the precise mathematical relations between PCI and empirical Φ measures at the whole-brain level?

Can we develop empirical Φ measures that are applicable to large, nonlinear systems and robust to experimental and methodological variations?

Can empirical Φ measures be extended in theoretically principled ways to address the neural bases of conscious contents?

How can ΦID be used to investigate, empirically, the relationship between consciousness and neural information dynamics?

At the same time, it is important to recognise that PID (and hence ΦID) are active areas of research and also face significant challenges. In particular, there are as yet no consensus formulae for the distinct information atoms (nor is it clear that there should be one) and calculating all the atoms remains intractable for large systems. However, recent research suggests that, for many systems, results might not be substantively impacted by the precise choice of PID [62] and in practical analyses different PID formulae often behave similarly [11,63]. Also, there are now various ways to coarse-grain PID to reduce its algorithmic complexity from super-exponential to linear [62,64] and separate work is making applications more robust via better estimators [65] and optimisation routines [66]. Altogether, PID is evolving from a purely mathematical construct to a widely applicable tool, thus offering weak IIT a powerful language not only to formulate, but also to reliably test hypotheses about consciousness in neural data.

Concluding remarks

At present, there is a tension in the development of IIT: ongoing work is developing the foundations of the theory without making it more empirically tractable, while the available experimental evidence does not lend specific support to the most distinctive claims of the theory. This tension is reflected in the opinions of the research community: a recent survey found that measures of neural integration were considered the most favoured indicators of consciousness, yet IIT was considered less promising than competing theories of consciousness [13]. To help resolve this tension and to allow both theoretical and empirical work inspired by IIT to flourish, we have drawn a pragmatic distinction between 'strong' and 'weak' versions of the theory.

Strong IIT is epitomised by the continued search for fundamental formulae relating phenomenology with physics, geared towards resolving the hard problem of consciousness [67]. By contrast, weak IIT takes its cue from the real problem [68] of developing tools to explain, predict, and control [69] features of consciousness, in this case via measures of information dynamics. Importantly, weak IIT remains motivated by theory but is closely guided by progressive empirical testing. In fact, the principles behind weak IIT have been implicitly incorporated into other theoretical frameworks [5], even if they have never, until now, been formalised and consolidated together.

Ultimately, we believe the IIT enterprise does hold enormous promise for advancing our scientific understanding of consciousness. Distinguishing these two flavours of IIT will, we hope, enable both lines of research to focus on the problems that matter (see Outstanding questions) and contribute to mutually beneficial exchanges and perhaps even, in the end, a satisfying convergence.

Acknowledgments

P.M. and D.B. are funded by the Wellcome Trust (grant no. 210920/Z/18/Z). A.B. and A.S. are supported by the Dr Mortimer and Theresa Sackler Foundation via the Sackler Centre for Consciousness Science; A.S. is additionally supported by the European Research Council (Advanced Investigator Grant 101019254) and by the Canadian Institute for Advanced Research (CIFAR Program on Brain, Mind, and Consciousness). F.R. is supported by the Ad Astra Chandaria foundation.

Declaration of interests

No interests are declared.

References

1. Tononi, G. *et al.* (2016) Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* 17, 450
2. Oizumi, M. *et al.* (2014) From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computat. Biol.* 10, e1003588
3. Massimini, M. *et al.* (2005) Breakdown of cortical effective connectivity during sleep. *Science* 309, 2228–2232
4. Casali, A.G. *et al.* (2013) A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* 5, 198ra105
5. Sarasso, S. *et al.* (2021) Consciousness and complexity: a consilience of evidence. *Neurosci. Conscious.* Published online August 30, 2021. https://doi.org/10.1093/nc/niab023
6. Re, V.L. *et al.* (2021) Role of transcranial magnetic stimulation (TMS) combined with electroencephalography (EEG) in disorders of consciousness (DOC). *J. Neurol. Sci.* 429, 118507
7. Casarotto, S. *et al.* (2016) Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann. Neurol.* 80, 718–729
8. Barrett, A.B. and Mediano, P.A. (2019) The Phi measure of integrated information is not well-defined for general physical systems. *J. Conscious. Stud.* 26, 11–20
9. Moon, K. (2019) Exclusion and underdetermined qualia. *Entropy* 21, 405
10. Dehaene, S. *et al.* (2014) Toward a computational theory of conscious processing. *Curr. Opin. Neurobiol.* 25, 76–84
11. Luppi, A.I. *et al.* (2020) A synergistic workspace for human consciousness revealed by integrated information decomposition. *bioRxiv* Published online November 26, 2020. https://doi.org/10.1101/2020.11.25.398081
12. Michel, M. *et al.* (2018) An informal internet survey on the current state of consciousness science. *Front. Psychol.* 9, 2134
13. Francken, J. *et al.* (2021) An academic survey on theoretical foundations, common assumptions and the current state of the field of consciousness science. *PsyArXiv* Published online June 14, 2021. https://doi.org/10.31234/osf.io/8mbsk
14. Merker, B. *et al.* (2021) The integrated information theory of consciousness: a case of mistaken identity. *Behav. Brain Sci.* 45, e41
15. Michel, M. and Lau, H. (2020) On the dangers of conflating strong and weak versions of a theory of consciousness. *PhiMiSci* Published online December 30, 2021. https://doi.org/10.33735/phimisci.2020.II.54
16. Doerig, A. *et al.* (2019) The unfolding argument: why IIT and other causal structure theories cannot explain consciousness. *Conscious. Cogn.* 72, 49–59
17. Melloni, L. *et al.* (2021) Making the hard problem of consciousness easier. *Science* 372, 911–912
18. Searle, J.R. (1980) Minds, brains, and programs. *Behav. Brain Sci.* 3, 417–424
19. Bedau, M. (2002) Downward causation and the autonomy of weak emergence. *Principia* 6, 5–50
20. Chalmers, D.J. (2006) Strong and weak emergence. In *The Re-Emergence of Emergence: The Emergentist Hypothesis From Science to Religion* (Davies, P. and Clayton, P., eds), pp. 244–256, Oxford University Press
21. Seth, A.K. (2010) Measuring autonomy and emergence via Granger causality. *Artif. Life* 16, 179–196
22. Seth, A.K. (2009) Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* 1, 50–63
23. Balduzzi, D. and Tononi, G. (2008) Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* 4, e1000091

24. Barbosa, L.S. *et al.* (2021) Mechanism integrated information. *Entropy* 23, 362
25. Tononi, G. and Koch, C. (2015) Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. B Biol. Sci.* 370, 20140167
26. Mørch, H.H. (2019) Is consciousness intrinsic?: A problem for the integrated information theory. *J. Conscious. Stud.* 26, 133–162
27. Fuliński, A. *et al.* (1998) Non-Markovian character of ionic current fluctuations in membrane channels. *Phys. Rev. E* 58, 919
28. Krohn, S. and Ostwald, D. (2017) Computing integrated information. *Neurosci. Conscious.* 2017, nix017
29. Kleiner, J. and Tull, S. (2021) The mathematical structure of integrated information theory. *Front. Appl. Math. Stat.* 6, 74
30. Barrett, A.B. (2014) An integration of integrated information theory with fundamental physics. *Front. Psychol.* 5, 63
31. Barbosa, L.S. *et al.* (2020) A measure for intrinsic information. *Sci. Rep.* 10, 1–9
32. Bayne, T. (2018) On the axiomatic foundations of the integrated information theory of consciousness. *Neurosci. Conscious.* 2018, niy007
33. Haun, A. and Tononi, G. (2019) Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy* 21, 1160
34. Lord, L.D. *et al.* (2017) Understanding principles of integration and segregation using whole-brain computational connectomics: implications for neuropsychiatric disorders. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 375, 20160283
35. Luppi, A.I. *et al.* (2019) Consciousness-specific dynamic interactions of brain integration and functional diversity. *Nat. Commun.* 10, 1–12
36. Canales-Johnson, A. *et al.* (2020) Dissociable neural information dynamics of perceptual integration and differentiation during bistable perception. *Cereb. Cortex* 30, 4563–4580
37. Ferrarelli, F. *et al.* (2010) Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proc. Natl. Acad. Sci.* 107, 2681–2686
38. Massimini, M. *et al.* (2009) A perturbational approach for evaluating the brain's capacity for consciousness. *Prog. Brain Res.* 177, 201–214
39. Nilsen, A.S. *et al.* (2019) Evaluating approximations and heuristic measures of integrated information. *Entropy* 21, 525
40. Barrett, A.B. and Seth, A.K. (2011) Practical measures of integrated information for time-series data. *PLoS Comput. Biol.* 7, e1001052
41. Oizumi, M. *et al.* (2016) Measuring integrated information from the decoding perspective. *PLoS Comput. Biol.* 12, e1004654
42. Barnett, L. and Seth, A.K. (2011) Behaviour of Granger causality under filtering: theoretical invariance and practical application. *J. Neurosci. Methods* 201, 404–419
43. Mediano, P.A. *et al.* (2021) Towards an extended taxonomy of information dynamics via integrated information decomposition. *arXiv* Published online September 27, 2021. https://doi.org/10.48550/arXiv.2109.13186
44. Tegmark, M. (2016) Improved measures of integrated information. *PLoS Comput. Biol.* 12, e1005123
45. Mediano, P.A. *et al.* (2019) Measuring integrated information: comparison of candidate measures in theory and simulation. *Entropy* 21, 17
46. Oizumi, M. *et al.* (2016) Unified framework for information integration based on information geometry. *Proc. Natl. Acad. Sci. U. S. A.* 113, 14817–14822
47. Langer, C. and Ay, N. (2020) Complexity as causal information integration. *Entropy* 22, 1107
48. Afrasiabi, M. *et al.* (2021) Consciousness depends on integration between parietal cortex, striatum, and thalamus. *Cell Syst.* 12, 363–373
49. Kim, H. and Lee, U. (2019) Criticality as a determinant of integrated information ϕ in human brain networks. *Entropy* 21, 981
50. Lee, U. *et al.* (2009) Propofol induction reduces the capacity for neural information integration: implications for the mechanism of consciousness and general anesthesia. *Conscious. Cogn.* 18, 56–64
51. Kim, H. *et al.* (2018) Estimating the integrated information measure Phi from high-density electroencephalography during states of consciousness in humans. *Front. Hum. Neurosci.* 12, 42
52. Shanahan, M. (2010) *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*, Oxford University Press
53. Bayne, T. and Carter, O. (2018) Dimensions of consciousness and the psychedelic state. *Neurosci. Conscious.* 2018, niy008
54. Birch, J. *et al.* (2020) Dimensions of animal consciousness. *Trends Cogn. Sci.* 24, 789–801
55. Fortier-Davy, M. and Millière, R. (2020) The multi-dimensional approach to drug-induced states: a commentary on Bayne and Carter's "dimensions of consciousness and the psychedelic state". *Neurosci. Conscious.* 2020, niaa004
56. Chennu, S. *et al.* (2014) Spectral signatures of reorganised brain networks in disorders of consciousness. *PLoS Comput. Biol.* 10, e1003887
57. Barnett, L. *et al.* (2020) Decreased directed functional connectivity in the psychedelic state. *NeuroImage* 209, 116462
58. Schartner, M.M. *et al.* (2017) Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Sci. Reports* 7, 46421
59. Mediano, P.A. *et al.* (2020) Effects of external stimulation on psychedelic state neurodynamics. *bioRxiv* Published online November 2, 2020. https://doi.org/10.1101/2020.11.01.356071
60. Luppi, A.I. *et al.* (2021) What it is like to be a bit: an integrated information decomposition account of emergent mental phenomena. *Neurosci. Conscious.* 2021, niab027
61. Williams, P.L. and Beer, R.D. (2010) Nonnegative decomposition of multivariate information. *arXiv* Published online April 14, 2020. https://doi.org/10.48550/arXiv.1004.2515
62. Rosas, F.E. *et al.* (2020) An operational information decomposition via synergistic disclosure. *J. Phys. A Math. Theor.* 53, 485001
63. Tax, T.M. *et al.* (2017) The partial information decomposition of generative neural network models. *Entropy* 19, 474
64. Rosas, F.E. *et al.* (2020) Reconciling emergences: an information-theoretic approach to identify causal emergence in multivariate data. *PLoS Comput. Biol.* 16, e1008289
65. Kleinman, M. *et al.* (2021) Redundant information neural estimation. *Entropy* 23, 922
66. Makkeh, A. *et al.* (2018) BROJA-2PID: A robust estimator for Bertschinger et al.'s bivariate partial information decomposition. *Entropy* 20, 271
67. Chalmers, D.J. (1995) Facing up to the problem of consciousness. *J. Conscious. Stud.* 2, 200–219
68. Seth, A.K. (2016) *The Real Problem*, Aeon
69. Seth, A. (2021) *Being You: A New Science of Consciousness*, Penguin
70. Tononi, G. and Edelman, G.M. (1998) Consciousness and complexity. *Science* 282, 1846–1851
71. Tononi, G. *et al.* (1998) Complexity and coherency: integrating information in the brain. *Trends Cogn. Sci.* 2, 474–484
72. Tononi, G. *et al.* (1994) A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad.Sci. U. S. A.* 91, 5033–5037
73. Tononi, G. and Sporns, O. (2003) Measuring information integration. *BMC Neurosci.* 4, 31
74. Tononi, G. (2004) An information integration theory of consciousness. *BMC Neurosci.* 5, 42
75. Balduzzi, D. and Tononi, G. (2009) Qualia: the geometry of integrated information. *PLoS Comput. Biol.* 5, 1–24
76. Bayne, T. (2010) *The Unity of Consciousness*, Oxford University Press
77. Rosas, F.E. *et al.* (2019) Quantifying high-order interdependencies via multivariate extensions of the mutual information. *Phys. Rev. E* 100, 032305