Check for updates

Theories of consciousness

Anil K. Seth[1,2]✉ and Tim Bayne[2,3,4]

Abstract | Recent years have seen a blossoming of theories about the biological and physical basis of consciousness. Good theories guide empirical research, allowing us to interpret data, develop new experimental techniques and expand our capacity to manipulate the phenomenon of interest. Indeed, it is only when couched in terms of a theory that empirical discoveries can ultimately deliver a satisfying understanding of a phenomenon. However, in the case of consciousness, it is unclear how current theories relate to each other, or whether they can be empirically distinguished. To clarify this complicated landscape, we review four prominent theoretical approaches to consciousness: higher-order theories, global workspace theories, re-entry and predictive processing theories and integrated information theory. We describe the key characteristics of each approach by identifying which aspects of consciousness they propose to explain, what their neurobiological commitments are and what empirical data are adduced in their support. We consider how some prominent empirical debates might distinguish among these theories, and we outline three ways in which theories need to be developed to deliver a mature regimen of theory-testing in the neuroscience of consciousness. There are good reasons to think that the iterative development, testing and comparison of theories of consciousness will lead to a deeper understanding of this most profound of mysteries.

Neural correlates of consciousness
(NCCs). The minimal set of neural events that is jointly sufficient for a conscious state.

[1]Department of Informatics and Sackler Centre for Consciousness Science, University of Sussex, Brighton, UK.

[2]Program on Brain, Mind, and Consciousness, Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada.

[3]School of Philosophical, Historical, and International Studies, Monash University, Melbourne, Victoria, Australia.

[4]Monash Centre for Consciousness and Contemplative Studies, Monash University, Melbourne, Victoria, Australia.

✉e-mail: a.k.seth@sussex.ac.uk

https://doi.org/10.1038/s41583-022-00587-4

In the early decades of its resurgence, the scientific study of consciousness focused on the search for the 'neural correlates of consciousness' (NCCs). Formally, the NCCs of a conscious state are the minimal set of neural events jointly sufficient for that state; in practice, the search for the NCCs has involved seeking the brain states and processes that are most closely related to consciousness[1–3]. Focusing on searching for the NCCs has been useful because the notion of NCCs is relatively 'theory neutral', and thus the NCC framework provided a common language and methodology for researchers with different theoretical and even metaphysical commitments. However, the limitations of the NCC framework have become increasingly clear, as revealed for example in the challenges involved in distinguishing 'true' NCCs from the neural prerequisites and consequences of consciousness[4–7]. In response to these limitations, there has been a steadily increasing focus on the development of theories of consciousness (ToCs). With a ToC in hand, we would be able to go beyond an NCC-based methodology and move towards models of consciousness that deliver explanatory insight. Indeed, having an empirically validated ToC should be the primary goal of consciousness science[8,9].

Whereas the NCC approach prioritizes the search for correlations between brain activity and consciousness, a theoretical approach instead focuses on identifying explanatory links between neural mechanisms and aspects of consciousness[10]. That being said, theorists often employ different conceptions of what it would take to secure an explanatory link between neural activity and consciousness. Some assume that a satisfactory ToC should and can close the 'explanatory gap' (BOX 1), and that it will be possible to render the relationship between neural activity and consciousness as transparent as the relationship between water's chemical structure and its gross behavioural profile[11]. Others doubt or remain agnostic as to whether the explanatory gap will ever be fully closed, but nonetheless hope for a framework that might explain certain aspects of consciousness and, in so doing, reduce or eliminate the sense of mystery surrounding its biophysical basis[12,13]. Still others argue that explanatory gap intuitions are misleading, and should not be taken seriously by the science of consciousness[14,15].

There is now a wide range of candidate ToCs (TABLE 1). Notably, instead of ToCs progressively being 'ruled out' as empirical data accumulate, they seem to be proliferating. This proliferation has led to both attempts to integrate existing theories with each other[16] and the development of 'adversarial collaborations', in which proponents of competing theories agree in advance about whether the outcome of a proposed experiment will support or undermine their preferred theory[17]. However, there are significant challenges to both theory integration and adversarial collaboration, as we discuss.

Box 1 | Theories of consciousness and the 'hard problem'

In the 1990s, David Chalmers famously distinguished between the 'hard' and 'easy' problems of consciousness[164]. The easy problems are concerned with the functions and behaviours associated with consciousness, whereas the hard problem concerns the experiential (phenomenal, subjective) dimensions of consciousness. What makes the hard problem hard is the 'explanatory gap'[165] — the intuition that there seems to be no prospect of a fully reductive explanation of experience in physical or functional terms.

Some theories of consciousness (ToCs) (for example, integrated information theory (IIT) and certain versions of higher-order theory (HOT)) address the hard problem directly. Other theories (for example, global workspace theories (GWTs)) focus on the functional and behavioural properties associated with consciousness; although they can be viewed as addressing the hard problem, this is not the primary goal of their proponents. A third strategy (adopted by some predictive processing theorists) aims to provide a framework in which various questions about the phenomenal properties of consciousness can be addressed, without attempting to account for the existence of phenomenology as such[67] — an approach sometimes called the 'real problem'[13,166].

A critical question in this area is whether the hard problem is indeed a genuine challenge that ought to be addressed by a science of consciousness, or whether it ought to be dissolved rather than solved. Those who take the latter view often argue that the appearance of a distinctively hard problem derives from the peculiar features of the concepts ('phenomenal concepts') that we employ in representing our own conscious states[167,168]. A related view is illusionism, according to which we do not actually have phenomenal states but merely represent ourselves as having such states[14,15]. Whatever the merits of these proposals, it seems likely that the grip of the hard problem may loosen as our capacity to explain, predict and control both phenomenological and functional properties of consciousness expands[166,169].

In this Review, we consider how a range of ToCs relate to each other and to empirical data, and we identify some promising avenues by which theory development and empirical research can jointly support each other in the search for a satisfying scientific account of conscious experience. Our attention is restricted to theories that are either themselves expressed in neurobiological terms or are plausibly taken to entail claims that can be expressed in neurobiological terms. (As we will see, some 'neurobiological' ToCs are expressed in the abstract language of functional relations or information theory, and qualify as 'neurobiological' only because the abstract features that they appeal to are associated with particular neural mechanisms.) We also consider only neuroscientific theories that are consistent with known physical theory, and we also leave to one side theories that link consciousness directly to quantum mechanical processes (for examples, see REFS[18,19]).

Preliminaries

One of the main reasons why ToCs 'talk' past each other is that they often have different explanatory targets. We therefore begin by considering what a comprehensive ToC should aim to account for, noting that even this issue is contested, with theorists often disagreeing about what kinds of phenomena a ToC should explain.

The heart of the problem of consciousness is the issue of 'experience' or 'subjective awareness'. Although no non-circular definition of these terms can be provided, the target phenomenon can be illuminated through some intuitive distinctions. There is 'something it is like' for an organism to be conscious[20], and what it is like to be in one state of consciousness differs from what it is like to be in another state of consciousness. A comprehensive ToC will explain why some organisms or systems are conscious whereas others are not, and it will also explain why states of consciousness differ from each other in the ways that they do.

States of consciousness can be grouped into two classes: global states and local states. Global states concern an organism's overall subjective profile and are associated with changes in arousal and behavioural responsiveness. Familiar global states include wakefulness, dreaming, sedation, the minimally conscious state, and (perhaps) the psychedelic state. These global states are sometimes called 'levels' of consciousness, but we prefer the term 'global states' because it leaves open the possibility that these states cannot be given a complete ordering in terms of a single dimension, but are best conceptualized as regions within a multidimensional space[21].

Local states — often referred to as 'conscious contents' or as states having 'qualia' — are characterized by 'what it is like' to be in them. The local state associated with having a headache is distinct from the local state associated with smelling coffee, for what it is like to have a headache differs from what it is like to smell coffee. Local states can be described at different levels of granularity, from low-level perceptual features (for example, colour), to objects, to complete multimodal perceptual scenes. An important subset of local states underpins the experience of selfhood, which encompasses experiences of mood, emotion, volition, body ownership, explicit autobiographical memory and the like[13,22–24]. Although neurobiological theories tend to focus on local states with sensory and perceptual content, consciousness also includes local states with cognitive and propositional content, such as the thoughts that arise when solving a crossword puzzle. Importantly, the local states that an agent has at a particular time do not simply occur as independent elements but are, instead, bound together as components of a single conscious scene that subsumes each of the agent's local states[25,26].

A second distinction is between the phenomenal properties of consciousness and its functional properties. The former term refers to the experiential character of consciousness, as is suggested by the phrase 'what it is like'. The functional aspects of consciousness concern the role(s) that mental states play in the cognitive economy of an organism in virtue of being conscious. ('Function' here encompasses both teleological functions — functional roles as shaped by evolution — and dispositional functions — the role a process plays in the operation of a larger system of which it is a part; see REF.[27]) For example, consciously seeing a coffee cup may enable a range of functions, such as the ability to behave flexibly with respect to the cup (perhaps to drink from it, or to throw it across the room), to lay down an episodic memory of the event, and to provide verbal reports about the experience. In making this distinction, we are not claiming that phenomenal and functional properties are independent (they are very likely not to be independent), merely that they provide distinct explanatory targets for ToCs. As we will see, some ToCs focus on the phenomenal features of consciousness, others focus on the functional features of consciousness and still others attempt to account for both the functional and phenomenal features of consciousness.

Explanatory gap intuitions
Intuitions that there is no prospect of a fully satisfying explanation of consciousness in physical, mechanistic terms.

Adversarial collaborations
Research projects in which proponents of different theories together design an experiment to distinguish their preferred theories, and agree in advance about how the outcome will favour one theory over the other(s).

Global states
Relating to an organism's overall state of consciousness, usually linked to arousal and behavioural responsiveness, and associated with the 'level' of consciousness.

Local states
Relating to particular conscious mental states, such as a conscious perception, emotion or thought. Local states are also often called conscious contents.

Binocular rivalry
A phenomenon in which different images are presented to each eye, and conscious perception alternates between the two images.

A third distinction is between two kinds of questions concerning local states ('contents') that a ToC might attempt to answer. On one hand, one might ask why an agent is in a certain local state (rather than another). On the other, one might ask why a particular local state has the experiential character that it has (rather than an experiential character of some other kind). This distinction can be explained with reference to binocular rivalry, in which each eye is presented with a different stimulus (say, a house to the right eye and a face to the left eye), and the subject's visual experience alternates between the left-eye stimulus and the right-eye stimulus[28]. Take a particular time at which the contents of consciousness involve a house, whereas the face is not consciously perceived. Here, we can ask why the mental state corresponding to 'house' is conscious (and that of 'face' is unconscious), and we can also ask why visual experiences of a house have the distinctive experiential character that they have rather than, say, the experiential character of seeing a face, hearing a bell or feeling pain.

Table 1 | **A selection of theories of consciousness**

Theory	Primary claim	Key refs
Higher-order theory (HOT)	Consciousness depends on meta-representations of lower-order mental states	[31,46]
Self-organizing meta-representational theory	Consciousness is the brain's (meta-representational) theory about itself	[34,140]
Attended intermediate representation theory	Consciousness depends on the attentional amplification of intermediate-level representations	[141,142]
Global workspace theories (GWTs)	Consciousness depends on ignition and broadcast within a neuronal global workspace where fronto-parietal cortical regions play a central, hub-like role	[47–49]
Integrated information theory (IIT)	Consciousness is identical to the cause–effect structure of a physical substrate that specifies a maximum of irreducible integrated information	[57,59,60]
Information closure theory	Consciousness depends on non-trivial information closure with respect to an environment at particular coarse-grained scales	[143]
Dynamic core theory	Consciousness depends on a functional cluster of neural activity combining high levels of dynamical integration and differentiation	[144]
Neural Darwinism	Consciousness depends on re-entrant interactions reflecting a history of value-dependent learning events shaped by selectionist principles	[145,146]
Local recurrency	Consciousness depends on local recurrent or re-entrant cortical processing and promotes learning	[65,71]
Predictive processing	Perception depends on predictive inference of the causes of sensory signals; provides a framework for systematically mapping neural mechanisms to aspects of consciousness	[67,73,79]
Neuro-representationalism	Consciousness depends on multilevel neurally encoded predictive representations	[84]
Active inference	Although views vary, in one version consciousness depends on temporally and counterfactually deep inference about self-generated actions	[76]; see also[91]
Beast machine theory	Consciousness is grounded in allostatic control-oriented predictive inference	[13,75,77]; see also[90]
Neural subjective frame	Consciousness depends on neural maps of the bodily state providing a first-person perspective	[24]
Self comes to mind theory	Consciousness depends on interactions between homeostatic routines and multilevel interoceptive maps, with affect and feeling at the core	[23,147]
Attention schema theory	Consciousness depends on a neurally encoded model of the control of attention	[148]
Multiple drafts model	Consciousness depends on multiple (potentially inconsistent) representations rather than a single, unified representation that is available to a central system	[149]
Sensorimotor theory	Consciousness depends on mastery of the laws governing sensorimotor contingencies	[88]
Unlimited associative learning	Consciousness depends on a form of learning which enables an organism to link motivational value with stimuli or actions that are novel, compound and non-reflex inducing	[150]
Dendritic integration theory	Consciousness depends on integration of top-down and bottom-up signalling at a cellular level	[151]
Electromagnetic field theory	Consciousness is identical to physically integrated, and causally active, information encoded in the brain's global electromagnetic field	[152]
Orchestrated objective reduction	Consciousness depends on quantum computations within microtubules inside neurons	[18]

Our selection of theories includes those that are either neurobiological in nature or potentially expressible in neurobiological terms.

Phenomenal character
The experiential nature of a local state, such as the 'redness' of an experience of red or the pain of a toothache — sometimes also called qualia.

Meta-representation
A mental representation that has as its target another mental representation

Notably, there may be some contents that cannot be conscious (for example, low-level processing within early sensory or regulatory systems) and others that can only be conscious (for example, globally integrated perceptual scenes). Thus, in addition to explaining why some mental contents are conscious in some contexts but not others, another challenge is to explain why some contents can never be conscious and why others can exist only as conscious.

Rather than address the full range of issues that we have just identified, most ToCs aim to explain only certain aspects of consciousness, perhaps as a step on the way to becoming comprehensive. Although being restricted in some way is not itself an objection to a ToC, it does mean that the task of inter-theory comparison is less straightforward than it might otherwise be. If theories are targeting different aspects of consciousness (say, one theory is focused on the phenomenal character of consciousness and another is focused on its functional profile) then they might not be the 'adversaries' that they at first glance appear to be.

The ToCs we review here are grouped into four categories: higher-order theories (HOTs), global workspace theories (GWTs), integrated information theory (IIT) and re-entry and predictive processing theories. Although some accounts of consciousness straddle multiple categories, and others are not plausibly subsumed under any of these categories (TABLE 1), this four-way distinction between ToCs provides a useful lens through which to view the current state of play in the science of consciousness (BOX 2; for other ways of grouping theories, see for example REF.[29]). In what follows, we introduce the key elements of each category, describe some notable within-category differences, and identify those aspects of consciousness most closely associated with each category. We then illustrate how these ToCs relate to each other in terms of some prominent empirical debates, and present several proposals that, we suggest, will help to drive a virtuous cycle between theory development and experimental investigation.

Box 2 | Other approaches: attention, learning and affect

The landscape of theories of consciousness (ToCs) includes numerous other theoretical approaches in addition to those surveyed in this Review (TABLE 1). One approach focuses on attention. For example, Graziano's attention schema theory associates conscious perception with a model of the control of attention[148]. Another attention-based ToC is the attended intermediate representational theory. First proposed by Jackendoff[141] and defended in detail by Prinz[142], this theory holds that consciousness occurs when intermediate-level perceptual representations gain access to attention.

Other theoretical approaches focus on learning. These include the proposal by Jablonka and Ginsburg that minimal consciousness is underpinned by a form of associative learning they term 'unlimited associative learning'. According to their proposal, this form of learning enables an organism to link motivational value with stimuli or actions that are novel, compound and non-reflex inducing[150]. Other learning-based theories overlap with some theories we have already described, such as Cleeremans' version of higher-order theory (HOT)[34,140] and Lamme's local recurrency account, which holds that recurrent signalling underpins consciousness in virtue of its role in learning[65]. Learning-based theories are also closely related to 'selectionist' approaches, which ground consciousness in evolutionary-like dynamics within and between neuronal populations[145,146].

Affect-based theories emphasize the brain's role in physiological regulation as the basis for consciousness. These theories include Damasio's proposal that consciousness depends on hierarchically nested representations of the organism's physiological condition[147,170], and proposals that mix an affect-based emphasis with predictive processing to ground conscious experiences in control-oriented interoceptive predictions[13,77,90]. Some affect-based theories deny that cortical mechanisms are necessary for consciousness, instead locating the mechanisms of consciousness in the brainstem[171,172] (although see REF.[173]).

Higher-order theories

The core claim that unites all HOTs is that a mental state is conscious in virtue of being the target of a certain kind of meta-representational state. Meta-representations are not merely representations that occur higher or deeper in a processing hierarchy but are, rather, representations that have as their targets other representations (FIG. 1). For example, a representation with the content 'I have a visual experience of a moving dot' is a meta-representation, for its content concerns the agent's own representations of the world rather than the world itself.

An important respect in which HOTs differ from each other concerns the account that they give of the nature and role of the meta-representations that are responsible for consciousness. Some versions of the approach identify the kinds of meta-representations that are crucial for consciousness with thoughts (or thought-like states) that have conceptual content[30–32]. Other varieties of HOT have been expressed in computational terms. According to the self-organizing meta-representational account, consciousness involves higher-order brain networks learning to redescribe the representations encoded in lower-order networks in a way that counts as meta-representational[33,34]. Alternatively, higher-order state space theory proposes that subjective reports (for example, statements such as 'I am aware of X') are metacognitive (higher-order) decisions about a generative model of perceptual content[35], whereas perceptual reality monitoring posits that conscious perception arises when a higher-order network judges a first-order representation to be a reliable reflection of the external world[36,37].

As should be clear from the foregoing, HOTs focus on explaining why some contents are conscious whereas others are not. However, these theories are not limited to this particular focus — they also have the resources to address issues pertaining to the experiential character of local states. One prominent example concerns the (debated) intuition that the contents of perceptual experience often outstrip the information available in 'first-order' sensory representations, as is alleged to occur in the context of peripheral vision[38,39]. The HOT-based proposal here is that the apparently 'inflated' phenomenology of peripheral visual experience is caused by the higher-order misrepresentation of first-order states[40]. The HOT approach can also be extended to explain why some contents are unable to be conscious (they cannot be the targets of appropriate meta-representational states) and why some contents are necessarily conscious (they are necessarily accompanied by appropriate meta-representational states). HOTs rarely focus on global states of consciousness, but it would be natural for them to appeal to the integrity of (meta-)representational processes to account for the distinctions between global states.

A particularly intriguing question is whether (and if so, how) HOTs explain the distinctive phenomenal character of various kinds of experiences. Why is the phenomenal character associated with seeing a sunset so different from the phenomenal character associated with a headache? The general shape of the higher-order response to this question is that the phenomenal character of a state is determined by the properties that the relevant meta-representational state ascribes to it. Most examples of this approach focus on visual experience[40], but there have also been higher-order attempts to account for the phenomenal character of emotional states[41] and metacognitive states, such as 'what it is like' to feel confident in a perceptual decision[42,43]. Ultimately, any fully reductive version of the higher-order approach must explain why the representation of various properties generates the phenomenology that it does (or is identical to it), and how neural activity enables the relevant properties to be represented in the first place.

Higher-order accounts of consciousness are primarily accounts of what makes a mental state conscious and, as such, the approach is not committed to any particular view of the function(s) of consciousness. Indeed, some HOTs downplay the idea that consciousness has any distinctive function[44]. Other versions of the higher-order approach identify the functional role of consciousness with the metacognitive processes associated with confidence judgements and error monitoring[45]. However, whereas higher-order views allow conscious mental states to be accompanied by conscious metacognitive judgements — such as those involved in explicit performance monitoring or subjective confidence reports — most versions of this approach do not require that conscious perception is always accompanied by a corresponding conscious metacognitive state. Instead, for meta-representations to be conscious, they themselves must be the objects of a suitable meta-representational state.

With respect to the neural basis of consciousness, the emphasis on meta-representation has led higher-order theorists to emphasize anterior cortical regions, especially the prefrontal cortex[30], given the association of these regions with complex cognitive functions. However, although most HOTs propose that anterior involvement is implicated in consciousness, there is disagreement about precisely which anterior regions (or processes) are required[46].

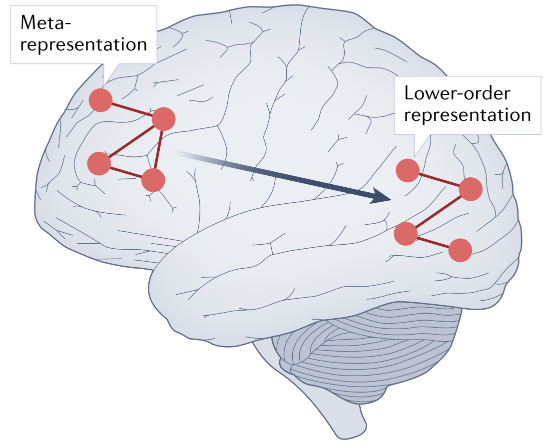

Fig. 1 | **Higher-order theories.** The core claim in higher-order theories (HOTs) of consciousness is that mental states are conscious in virtue of being the target of specific kinds of meta-representation. For example, lower-order representations of visual signals in posterior cortex would support conscious visual perception when targeted by the right kind of higher-order meta-representation. Supportive evidence for HOTs comes from studies implicating anterior cortical areas in conscious contents, with an emphasis on prefrontal cortex — especially when performance is matched across conscious and non-conscious conditions[30,100]. HOTs are also indirectly supported by lesion evidence linking metacognition to prefrontal areas[153]. These theories are challenged by evidence suggesting that anterior areas are not involved in consciousness[108,154], perhaps instead being necessary only for enabling subjective report and executive control[6]. Figure adapted with permission from REF.[46], Elsevier.

Global workspace theories

GWTs originate from 'blackboard' architectures in artificial intelligence, in which the blackboard is a centralized resource through which specialized processors share and receive information. The first GWT of consciousness[47] was framed at a cognitive level. It proposed that conscious mental states are those that are 'globally available' to a wide range of cognitive processes including attention, evaluation, memory and verbal report. The core claim of GWTs is that it is the wide accessibility of information to such consumer cognitive systems that constitutes conscious experience (FIG. 2).

This basic claim has since been developed into a neural theory — often referred to as the 'global neuronal workspace theory' — according to which sensory information gains access to consciousness when it is 'broadcast' within an anatomically widespread neuronal workspace that is implemented across higher-order cortical association areas, with a particular (although not exclusive) emphasis on the prefrontal cortex[48,49]. Access to the global workspace is achieved through nonlinear network 'ignition' in which recurrent processing amplifies and sustains neuronal representations[50]. The emphasis on ignition and broadcast — as compared with meta-representation — is one way in which GWT is distinguished from the HOT approach.

Like HOTs, GWTs focus on the question of what makes a representation conscious, and GWT theorists have rarely attempted to account for the phenomenal differences between distinct kinds of experiences (although see REF.[51]). Returning to our example of binocular rivalry, the GWT view aims to explain why, at a particular point in time, the mental state corresponding to 'house' is conscious (whereas that corresponding to 'face' is unconscious), but offers no direct account of the experiential contrast between seeing a house on the one hand and seeing a face on the other.

The relative silence of GWTs on the issue of experiential character aligns with the general tendency of such theories to focus on functional, rather than phenomenal, aspects of consciousness. In fact, GWTs are often explicitly proposed as accounts of 'conscious access'[49]; that is, as accounts of why certain representations are available to be flexibly used by a wide range of consuming systems

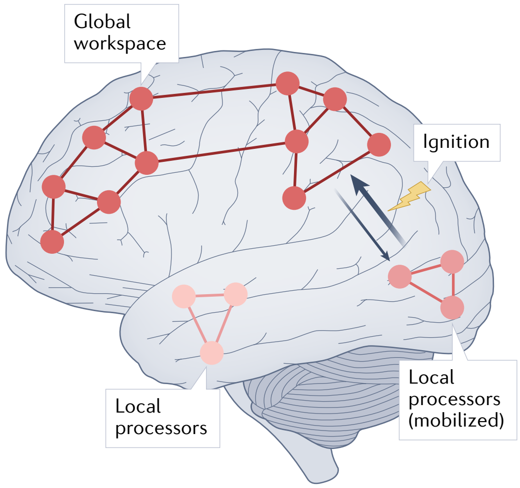

Fig. 2 | **Global workspace theories.** The core claim of global workspace theories (GWTs) of consciousness is that mental states are conscious when they are broadcast within a global workspace in which fronto-parietal networks play a central hub-like role. Activity in local processors (for example, sensory regions) becomes temporarily 'mobilized' into the workspace upon ignition[155]. Empirical support for GWTs comes from studies that have associated consciousness with neuronal signatures of ignition and long-distance information sharing[48,49,53,101]. Neural signatures of ignition are suggested by divergences of brain activity in anterior cortical regions at around 200–300 ms after stimulus onset, corresponding to trials with and without conscious perception[48,101], including in 'no-report paradigms'[111] (see also REF.[156]). Such studies have been recently extended to decoding; for example, activity patterns at around 300 ms after a stimulus predicted subjective reports in ways that generalized across sensory modalities[119]. Signatures of long-distance information for conscious versus unconscious content have been identified using a range of methods[49,102]. As with higher-order theories, GWTs are challenged by evidence that anterior regions might be involved in behavioural report rather than consciousness per se.

(whereas others are not). The core functional property addressed by GWTs is the ability of conscious states to guide behaviour and cognition in flexible, context-dependent ways. GWTs also offer clear accounts of how consciousness is related to other cognitive processes, such as attention and working memory. According to GWTs, attention selects and amplifies specific signals, allowing them to enter the workspace (and thus be conscious), whereas consciousness and working memory are intimately related because attended working memory items are conscious and use the global workspace for broadcast[49].

GWTs account for changes in global states of consciousness in terms of alterations to the functional integrity of the workspace. Neurally, a global loss of consciousness is reflected in impaired functional or dynamical connectivity in fronto-parietal regions that are considered 'hub' nodes in the global workspace[52], and in functional connectivity becoming increasingly constrained to patterns directly reflecting the underlying structural connectivity[53–55].

One important question raised by GWTs concerns what exactly is required for a workspace to qualify as 'global'[25,56]. Is it the number (and type) of consuming systems to which the workspace can broadcast that matters, or is it the kind of broadcasting that occurs within the workspace? Or are both of these considerations relevant to what counts as a 'global workspace'? These questions need to be answered if we are to know what predictions GWTs make with respect to consciousness in, for example, infants, individuals with brain damage, people who have undergone split-brain surgery, non-human animals and artificial intelligence systems.

Integrated information theory

IIT starts from a very different place from HOTs or GWTs by advancing a mathematical approach to characterizing phenomenology. The theory starts by proposing axioms about the phenomenological character of conscious experiences (that is, properties that are taken to be self-evidently true and to apply to all possible forms of consciousness), and from these axioms it derives claims about the properties that any physical substrate of consciousness must satisfy. IIT then proposes that physical systems that instantiate these properties necessarily also instantiate consciousness[57–60] (see FIG. 3). Specifically, IIT proposes that consciousness should be understood in terms of 'cause–effect power' associated with irreducible maxima of integrated information generated by a physical system. Integrated information, in turn, is associated with the information theoretic quantity Φ, which measures — broadly speaking — how much information is generated by a system as a whole, compared with its parts considered independently. In IIT, consciousness is an intrinsic, fundamental property of a system, and is determined both by the nature of the causal mechanisms that compose it and by their state[60].

In contrast to HOTs and GWTs, IIT links consciousness primarily with posterior cortical areas (the so-called 'posterior hot zone' encompassing parietal, temporal and occipital areas), in part on the grounds that these areas exhibit neuroanatomical properties that are supposedly well suited for generating high levels of integrated information[59]. Also in contrast to GWTs and HOTs, which associate consciousness with aspects of cortical information processing (that is, functional descriptions of what a system does), IIT does not refer to 'information processing' per se. Instead, it links consciousness to properties of the intrinsic cause–effect structure of a system; namely, to the causal power of a system to influence itself. According to IIT, any system that generates a non-zero maximum of (irreducible) integrated information is conscious, at least to some degree. Because of this, IIT would appear to imply that there already exist non-biological systems that are conscious[61].

IIT is reasonably comprehensive, offering accounts of both global states and local states of consciousness[59] (see FIG. 3). Global states are associated with the quantity of irreducible integrated information generated by a system, as measured by Φ. IIT therefore encourages a unidimensional conception of global states, for it equates an organism's level of consciousness with its value of Φ. The experiential character of local states can be understood

No-report paradigms
Behavioural experiments in which participants do not provide subjective (verbal, behavioural) reports.

Φ
The amount of information specified by a system that is irreducible to that specified by its parts. There are many variations of Φ, each calculated differently and making different assumptions.

Posterior hot zone
A range of brain regions towards the rear of the cortex, including parietal, temporal and occipital areas, as well as regions such as the precuneus.

Complex
In integrated information theory (IIT), a subset of a physical system that underpins a maximum of irreducible integrated information.

in terms of 'conceptual structures', which IIT treats as 'shapes' in a high-dimensional space that is specified by the mechanistic cause–effect structure of the system. These shapes underpin (or are identical to) specific kinds of phenomenal character. For example, the spatial nature of visual experience has been related to the cause–effect structure specified by grid-like mechanisms present in early visual cortex[62]. The global unity of consciousness is explained by the integrated aspect of integrated information — its association with information generated by the 'whole' over and above that generated by the 'parts'. Finally, according to IIT, contents are conscious (rather than unconscious) when, and only when, they are incorporated into a cause–effect 'complex' (where a complex is a subset of the physical system that underpins a maximum of irreducible integrated information).

Returning to the binocular rivalry example, IIT explains why the subject reports experiencing a house (rather than a face) by appealing to the hypothesis that the complex underlying their report is associated with the conceptual structure corresponding to the content 'house' (rather than the content 'face'), and it explains the experiential contrast between seeing a house and seeing a face in terms of the 'shape' of the corresponding conceptual structure.

Although IIT provides a more comprehensive treatment of the various aspects of consciousness than most ToCs, it says comparatively little about how consciousness is related to other aspects of the mind, such as attention, learning and memory, and has not yet focused on the relevance of embodiment and environmental embeddedness for consciousness (the latter also being a challenge for HOTs and GWTs). That said, IIT theorists have made initial steps towards addressing some of these challenges by, for example, developing measures of 'matching complexity' that track the shared information between an agent and its environment, and by formulating agent-based models in which agents that are able to engage effectively with their surroundings are found to exhibit increased amounts of integrated information[62–64].

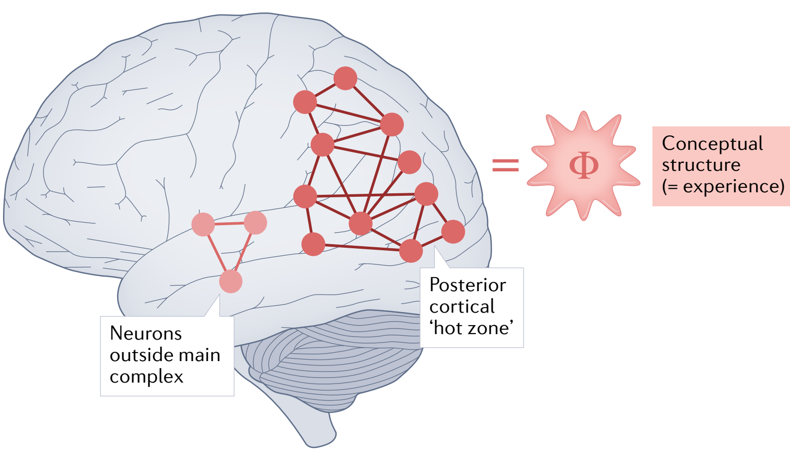

Fig. 3 | **Integrated information theory.** The core claim of integrated information theory (IIT) is that consciousness is identical to the cause–effect structure of a physical system that specifies a maximum of irreducible integrated information. The content of consciousness is associated with the form of the cause–effect structure, and the level of consciousness with its irreducibility, as measured by quantity Φ[59]. Anatomically, IIT is associated with a posterior cortical 'hot zone'. Empirical assessment of this core claim is challenging, largely because Φ is infeasible to measure, except in simple model systems. Various proxies for Φ have been developed[157] and some show promise. Prominent among them is the perturbational complexity index (PCI), which measures the algorithmic (Lempel–Ziv) complexity of brain responses to transcranial magnetic stimulation[158]. Importantly, the PCI has diagnostic and prognostic value in tracking global states of consciousness in patients with neurological impairments[158]. However, the PCI is not equivalent to Φ and correlations between PCI values and global states of consciousness are not incompatible with other theories of consciousness. Other evidence that indirectly supports IIT comes from psychophysical studies suggesting that local changes in the strength of lateral connections within visual cortex can alter the structure of visual space[116], and by evidence relating changes in global states to reduced functional diversity and integrative capacity in posterior cortical regions[159]. IIT would be challenged by evidence which indicates that activity in anterior cortical regions is necessary for perceptual consciousness.

Re-entry and predictive processing theories

Finally, we consider two general approaches to understanding consciousness that emphasize the importance of top-down signalling in shaping and enabling conscious perception. The first — re-entry theories — are ToCs as such, and associate conscious perception with top-down (recurrent, re-entrant) signalling[65,66]. The second group — predictive processing theories — are not, first and foremost, ToCs but are more general accounts of brain (and body) function that can be used to formulate explanations and predictions regarding properties of consciousness[67].

Re-entry theories are motivated by neurophysiological evidence revealing the importance of top-down signalling for conscious (usually visual) perception (for examples, see REFS[68–70]). In one prominent re-entry theory — local recurrency theory — Lamme argues that localized recurrent or re-entrant processing within perceptual cortices is sufficient to give rise to consciousness (given the integrity of other enabling factors, such as brainstem arousal), but that parietal and frontal regions might be required for reporting the contents of perceptual experience or drawing on them for reasoning and decision-making[65,71] (FIG. 4).

Broadly speaking, predictive processing theories have two motivations. One motivation traces to considering the problem of perception as one of inference about the causes of sensory signals[72,73]. The other — exemplified by the free energy principle[74] — appeals to fundamental constraints regarding control and regulation that apply to all systems that maintain their organization over time[75–77] (but see REF.[78]). Both lead to the notion that the brain implements a process of 'prediction error minimization'[79] that approximates Bayesian inference through the reciprocal exchange of (usually top-down) perceptual predictions and (usually bottom-up) prediction errors[80] (although see REF.[81]). Some expressions of predictive processing, such as active inference, add the notion that sensory prediction errors can be minimized not only by updating predictions but also by performing actions to bring about expected sensory data — thereby enabling a form of predictive control[82,83].

Although predictive processing theories did not originate as ToCs, it has been suggested that they can furnish systematic correlations between neural mechanisms and phenomenological properties[67], where 'systematic' means having explanatory power guided by theoretical

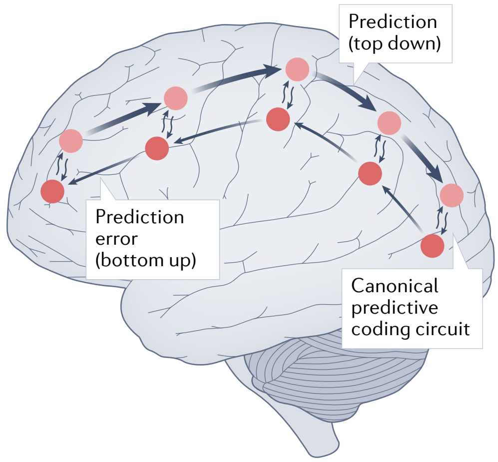

Fig. 4 | **Re-entry theory and predictive processing.** The core claim of re-entry theory and predictive processing is that conscious mental states are associated with top-down signalling (re-entry; thick arrows) that, for predictive processing, convey predictions about causes of sensory signals (thin arrows signify bottom-up prediction errors), so that continuous minimization of prediction errors implements an approximation to Bayesian inference. Conscious contents are specified — in most predictive processing theories — by the content of top-down predictions. Evidence in favour of these theories comes from studies that link top-down signalling with perceptual experience[68–70,160]. In further support of predictive processing, there is abundant evidence that expectations shape both the content of, and speed of access to, conscious perception[161–163], which some studies relate directly to top-down signalling[160]. These theories would be challenged by evidence that top-down signalling or predictive processing occurs in the absence of consciousness, or that changes in these processes do not affect conscious states.

considerations, in contrast to mere empirical correlations as in the vanilla NCC approach. From this perspective, predictive processing theories fulfil many of the desiderata for ToCs we outlined earlier, but may be best thought of as theories for consciousness science, rather than ToCs per se, for there are many perspectives on precisely how predictive processing relates to consciousness[84,85].

Predictive processing theories typically address local conscious states in terms of the content of top-down perceptual predictions[73,79,86,87]: informally, perceptual content is given by the brain's 'best guess' of the causes of its sensorium. The experiential character of a local state is specified by the nature of the perceptual predictions at play. For example, the phenomenology of 'objecthood' in vision may be accounted for by conditional predictions about the sensory consequences of actions[87,88], whereas the phenomenology of emotional states may be explained by the role of interoceptive predictions in regulating the organism's physiological condition[89,90]. The example of emotion highlights that predictive processing theories, more than the others discussed here, encompass issues related to conscious selfhood[13,77,91].

Predictive processing can explain the distinction between conscious and unconscious states in terms of whether a mental state is part of a current 'best guess' (or optimal posterior) during perceptual inference. In the example of binocular rivalry, predictive processing envisages two competing perceptual hypotheses (best guesses), one of which 'wins', leading to perceptual dominance. Sensory signals from the alternative hypothesis accumulate as prediction error, which eventually leads to a perceptual transition, at which point sensory signals explained by the previously dominant best guess now become a source of unexplained prediction error, and so the cycle repeats[92,93]. (The experiential contrast between the house and the face would, as mentioned, be explained by properties of the corresponding perceptual predictions.) In those varieties of predictive processing that emphasize active inference, a change in conscious content can happen only if perceptual belief updating comes about through action (where action can be overt, such as a saccadic eye movement, or covert, such as a shift of attentional focus)[76,94].

Predictive processing theories do not generally deal with global states of consciousness, but it would be natural for them to appeal to the integrity of the relevant predictive processes in explaining distinctions among global states[95], in much the same way in which HOT accounts can appeal to the integrity of the relevant meta-representational machinery.

With respect to the functional dimensions of consciousness, both re-entry and predictive processing approaches provide clear treatments of the relationship between consciousness and attention. In local recurrency theory, as in GWTs, attention provides a selective boost to sensory signals so that they reach prefrontal and parietal regions, engaging conscious access[71]. In predictive processing, attention is associated with the process of 'precision weighting', in which the estimated precision of sensory signals is modulated in ways intuitively equivalent to altering the signal-to-noise ratio or 'gain' of these signals[74,96]; and in active inference, as mentioned, attentional sampling may be necessary for changes in conscious content[76,94].

Evaluating theories of consciousness

The range of data that have been appealed to in connection with the debate between rival ToCs is vast, and we cannot hope to provide a full inventory of them here. Instead, we offer a selective overview of some current debates, highlighting the diversity of data that can be brought to bear on the evaluation of ToCs. (Some other empirical data generally used in support of each ToC are described in the legends of FIGS 1–4).

As a background point, it is important to recognize the holistic nature of theory evaluation. Theories are not confirmed by a single finding; nor are they generally defeated by a single experiment. Instead, theory confirmation is typically an incremental process, in which one theory wins out over its rivals by providing an account of the target phenomenon that explains a wide range of data and can be integrated with successful theories in neighbouring domains[97–99].

One obvious source of constraints on a ToC is the structure of consciousness. Although numerous structural features have been discussed in connection with ToCs, one structural feature of particular utility for contrasting ToCs is the unity of consciousness — the fact that

Interoceptive predictions
Predictions about the causes of sensory signals originating from within the body (interoception refers to perception of the body 'from within').

Unity of consciousness
The fact that that the experiences that a single agent has at a time seem always to occur as the components of a single complex experience.

Cognitive access
A functional property whereby a mental state has access to a wide range of cognitive processes, usually including verbal and/or behavioural report.

the experiences that a single agent has at a time seem always to occur as the components of a single complex experience, one that fully captures what it is like to be that agent[25]. Different ToCs take very different attitudes to the unity of consciousness. IIT places considerable emphasis on the unity of consciousness. It not only assumes that consciousness is always unified but also appeals to the claim that consciousness is necessarily unified to motivate the association of consciousness with (maxima of) irreducible integrated information. Although GWTs do not emphasize the unity of consciousness in the way that IIT does, the association of consciousness with broadcast within a functionally integrated workspace suggests that they too may have the resources to provide a plausible account of the unity of consciousness. Other ToCs, such as HOTs and re-entry/predictive processing theories, have a more ambivalent relationship to the unity of consciousness, tending either to only gesture at an account of this property or to overlook it entirely. The contrast in attitudes among ToCs to the unity of consciousness is due, in part at least, to more fundamental disagreement over whether consciousness is (necessarily) unified. Although the unity of consciousness promises to provide an important constraint on ToCs, in order for this promise to be realized we need a better account of the respects in which consciousness is (necessarily) unified.

A second source of constraints is provided by neural data. For example, it is generally accepted that the cerebellum is neither necessary nor sufficient for consciousness. A ToC ought to account for this fact, and explain why the cerebellum is not implicated in consciousness. Some ToCs readily provide such an explanation — for example, IIT argues that the cerebellum is not implicated in consciousness because its architecture is poorly suited for generating high levels of integrated information[59]. But explanations of this sort lend specific support to a theory only if the account provided is more plausible than the accounts that might be provided by its competitors, and whether that condition is satisfied is currently an open question. For example, advocates of HOTs could argue that the cerebellum lacks the capacity to support meta-representations of the relevant kind; proponents of GWTs can make the case that the cerebellum does not implement a global workspace; and re-entrant and predictive processing theorists can point to the absence of rich recurrent signalling in the cerebellum[65].

Although it is generally accepted that a ToC should explain why the cerebellum is not implicated in consciousness, there are other kinds of neural data that are much more controversial from the point of view of ToCs. An important example is provided by the debate about the role of prefrontal ('front-of-the-brain') processes in consciousness.

Using various experimental paradigms, many neuroimaging studies have found prefrontal engagement for conscious (versus unconscious) perception[48], based both on regional activity[100,101] and on functional connectivity between frontal and other regions[102]. A small number of primate studies have also found that conscious contents can be decoded from prefrontal activity patterns during binocular rivalry, continuous flash suppression and rapid serial presentation of visual stimuli[103–105]; see also REF.[106] for a more complex picture in which content-relevant information was decoded from a wide range of both activated and deactivated cortical regions during an object recognition task. Lesion evidence, and evidence from brain stimulation, has also been used to argue that the prefrontal activity is crucially implicated in consciousness (see REF.[30] for a review).

Advocates of HOTs and GWTs appeal to these findings to support their accounts over competing theories. In response, advocates of IIT and re-entry theories argue that the observed prefrontal activity is a (non-necessary) consequence of consciousness and is probably associated with cognitive access to the contents of consciousness and the ability to provide behavioural reports, rather than with conscious perception per se[107,108] (but see REF.[109]). Those who defend this 'back-of-the-brain' perspective argue that posterior cortical processes — encompassing parts of the perceptual and parietal cortex and precuneus — suffice for perceptual experience, and that 'front-of-the-brain' processes are not necessary. This claim is supported by so-called 'no-report' studies which have tended to find diminished prefrontal engagement when subjects do not provide explicit reports about their perceptions[6,110] (but see REF.[111]). 'Back-of-the-brain' advocates also draw on positive evidence in favour of a tight coupling between posterior activity and consciousness. For example, one innovative study that probed for conscious contents during sleep using a serial awakening paradigm found that activity in posterior cortical regions predicted whether an individual would report dream experience, across both rapid eye movement and non-rapid eye movement sleep stages[112] (but see REF.[113]). Finally, the 'front-of-the-brain' interpretation of decoding studies is open to challenge, for showing that conscious contents can be 'read out' from a particular area does not establish that the brain itself is 'reading out' those contents from that area in a way that constitutes a relevant kind of meta-representation or global broadcast.

Although some aspects of the 'front-of-the-brain' versus 'back-of-the-brain' debate do indeed concern neurobiological data —for example, opinions differ on where the anatomical boundaries of the prefrontal cortex lie[107,109] — at its heart is a disagreement about the relationship between consciousness and cognitive access: is it reasonable to take the availability of content for verbal report and the direct control of behaviour as a proxy for consciousness, or should investigations into the brain basis of consciousness remain neutral as to how exactly consciousness and cognitive access are related[114] (see also BOX 3). Debate about this question is reflected in the attitudes that different ToCs take to cognitive access. GWTs place cognitive access at the heart of their account of consciousness, suggesting not only that the contents of consciousness are always available for cognitive access but also that the processes that underlie cognitive access (namely, ignition and global broadcast) serve as the basis of conscious experience (see REF.[111] for a recent nuance on this view). Other theories, such as IIT and local recurrency accounts, deny a close relationship between consciousness and cognitive access, holding that mental states can be conscious

Box 3 | The measurement problem

To test a theory of consciousness (ToC), we need to be able to reliably detect both consciousness and its absence. At present, experimenters typically rely on a subject's introspective capacities, either directly or indirectly, to identify their states of consciousness. However, this approach is problematic, for not only is the reliability of introspection questionable, but there are many organisms or systems (for example, infants, individuals with brain damage and non-human animals) who might be conscious but are unable to produce introspective reports. Thus, there is a pressing need to identify non-introspective 'markers' or 'signatures' of consciousness.

Numerous such indicators have been proposed in recent years. Some of these — such as the perturbational complexity index (PCI)[158] — have been proposed as markers of consciousness as such, whereas others — such as the optokinetic nystagmus response[174] or distinctive bifurcations in neural dynamics[111] — have been proposed as markers of specific kinds of conscious contents. The former have been applied fruitfully to assessing global states of consciousness in individuals with brain injury[175] whereas the latter have been deployed in 'no-report' studies of conscious content, in which overt behavioural reports are not made[6]. Whatever its focus, however, any proposed indicator of consciousness must be validated: we need to know that it is both sensitive and specific. Although approaches to validation based on introspection have the problems mentioned above, theory-based approaches are also problematic. Because ToCs are themselves contentious, it seems unlikely that appealing to theory-based considerations could provide the kind of intersubjective validation required for an objective marker of consciousness. Solving the measurement problem thus seems to require a method of validation that is based neither solely on introspection nor on theoretical considerations. The literature contains a number of proposals for addressing this problem[114,176], but none is uncontroversial[177,178].

without being available for the direct control of thought and action, and also that mental states could, in principle, be available for the direct control of thought and action without being conscious. Although higher-order approaches are not committed to any particular relationship between consciousness and cognitive access, in practice their advocates generally assume that the contents of consciousness will be cognitively accessible (for example, see REF.[46]), although perhaps not vice versa.

Perhaps the most powerful source of data for evaluating rival ToCs involves novel predictions. Many of the most significant events in the history of science have involved the confirmation of novel predictions[115]. For example, general relativity received strong support from the fact that it predicted both the advance of the perihelion of Mercury and the way in which starlight grazing the Sun's surface would be deflected. If a ToC was to make confirmed novel predictions, then it would be strongly supported, especially when compared with theories that failed to make the relevant prediction, or made different and incompatible predictions.

Many of the novel predictions that contemporary ToCs make are difficult to test. For example, both the re-entry and IIT accounts predict that posterior cortical activity can support conscious experience without contribution from anterior areas, but at present we lack reliable methods to verify such claims, as verification relies on subjective report (or, at least, executive control in some form), which in turn requires anterior cortical activity. More dramatically, IIT predicts that consciousness is widely distributed throughout nature, including in many non-biological systems, and might even occur in systems that are as simple as photodiodes and single atoms (although, interestingly, not in strictly feedforward neural networks[61]). This prediction runs counter to widely held assumptions about the distribution of consciousness, but cannot be sensibly evaluated in the absence of robust methods for detecting the presence of consciousness in such systems (BOX 3).

In some cases, methodological advances may bring novel predictions within reach of testability. One striking prediction, arising from IIT, is that changes in neural structure could lead to changes in conscious experience even when these changes do not give rise to changes in neural activity[116]. For example, inactive neurons in the visual cortex may contribute to visual experience, whereas inactivated neurons would not[59]. This prediction arises because, in IIT, it is the cause–effect structure specified by neural mechanisms that matters for consciousness. This means that if one intervenes in neural mechanisms, so as to change the cause–effect structure, then consciousness can change even if the corresponding neural dynamics do not change — a prediction that is particularly counter-intuitive in the case where dynamics are absent (that is, for inactive neurons). Hypotheses such as this, which do not readily follow from the other theories discussed here, may be testable using precise interventional methods, such as optogenetics, in animal models of perceptual decision-making[117].

A particularly fruitful avenue for evaluating rival ToCs focuses on the temporal profile of conscious (as opposed to unconscious) processing, as reflected for example by event-related potentials in electrophysiological recordings. Some theorists (for example, see REF.[118]) argue that conscious perception has an early (120–200 ms) onset following stimulus presentation, appealing to evidence that suggests a robust correlation between perceptual consciousness and early-onset modality-specific negative-going event-related potentials — called awareness negativity responses — while questioning the reliability of previously discussed later-onset signatures, such as the P3b (a positive-going event-related potential observed at ~300 ms after stimulus onset). The early negativity highlighted by Dembski and colleagues has been found in both vision and audition, leading them to argue that there is a generalized early-onset response that robustly indexes perceptual consciousness. Such data point in favour of IIT and local re-entry accounts of consciousness (but see REF.[119] for a later cross-modal signature of conscious perception). Other theorists[120,121] argue in favour of a much later onset (roughly, 250–400 ms) for perceptual consciousness. Besides the debated P3b, late-onset accounts are motivated by various perceptual phenomena that appear to match this timescale, including the psychological refractory period, the attentional blink and postdictive effects — the last of these being of particular interest in showing that a delayed cue can retrospectively trigger conscious perception[122]. Candidate late-onset neural signatures of conscious perception include long-distance information sharing and bifurcation dynamics[49,111]. Evidence in favour of late-onset accounts of perceptual consciousness generally supports higher-order and global workspace ToCs. The debate between 'early-onset' and 'late-onset' accounts of perceptual experience is likely to remain a central topic of discussion for the foreseeable future. Note that the issue of the temporal profile of conscious processing is distinct from both the perception of duration[123] and the temporal characteristics of a conscious 'moment'[124,125], both of which reflect aspects of conscious content that ought to be explained by a ToC.

Moving forward

At present, ToCs are generally used as 'narrative structures' within the science of consciousness. Although they inform the interpretation of neural and behavioural data, it is still rare for a study to be designed with questions of theory validation in mind[126]. Although there is nothing wrong with employing theories in this manner, future progress will depend on experiments that enable ToCs to be tested and disambiguated. We conclude our review by identifying three issues that need to be addressed for a mature regimen of theory-testing to flourish in consciousness science.

First, ToCs need to be developed with precision, for theories that appeal only to vague and imprecise constructs can generate only vague and imprecise predictions. For example, HOTs and predictive processing and re-entry theories need to specify the kinds of meta-representations and re-entrant or predictive processes that are distinctive of (specific aspects of) consciousness; IIT needs to make precise its implications for the functional profile of consciousness and the impact of the environment and embodiment on consciousness; and GWTs need to provide a principled account of which workspaces qualify as 'global' in the relevant sense.

A promising approach here is to use computational models to bring mechanistic specificity to ToCs that may have been formulated in relatively abstract or conceptual terms. In addition to grounding the generation of fine-grained predictions, such models might also provide a shared language in which the relative merits of rival ToCs can be compared, which can be especially useful for comparing ToCs originating from different starting points. For example, computational models could reveal shared principles of top-down signalling among HOTs and re-entry and predictive processing theories, while clarifying the distinctions between meta-representation (for example, see REF.[35]) and global broadcast (for example, see REFS[127,128]) that separate HOTs from GWTs[129]. The development of computational models might also allow contrasts between ToCs to be reframed in terms of (potentially distributed) processes rather than, as is currently popular, in terms of broad neuroanatomical regions (for example, as in the debate between 'front-of-the-brain' and 'back-of-the-brain' theorists[111]). A key challenge for the computational approach is to develop models that do not merely account for the functional features of consciousness but also account for its phenomenological properties — a challenge that can be described by the general label of 'computational (neuro)phenomenology' (for examples, see REFS[37,130]). This brings the additional challenge of how to validate, or disambiguate between, computational models using phenomenological data (for example, see REF.[131]) — a challenge that can be met, at least in part, by collecting subjective reports at the appropriate levels of phenomenological granularity (BOX 3).

In addition to being made more specific, ToCs also need to be made more comprehensive. For the most part, ToCs have tended to focus on particular kinds of local states (perceptual experiences, with an emphasis on vision), on particular kinds of global states (ordinary waking awareness) and on particular kinds of conscious creatures (adult human beings). Although there are good reasons why theorists have tended to focus on a restricted class of conscious states and creatures — experimental accessibility being an important factor — a fully comprehensive ToC must do justice to the rich diversity of consciousness. With respect to local states, ToCs must go beyond perception and account also, for example, for affect, temporality, volition and thought. With respect to global states, ToCs must go beyond ordinary wakefulness and account also for the distinctive modes of consciousness associated with, for example, dreaming, meditation, disorders of consciousness and the psychedelic state. With respect to conscious creatures, ToCs must go beyond adult experience and address questions regarding consciousness in human infants, non-human animals and even artificial systems. Although there is nothing wrong with ToCs that have a restricted focus, theories that provide a more comprehensive account of consciousness have obvious advantages over those that do not, especially if they can identify explanatory connections between different aspects of consciousness.

The third issue to be addressed is the measurement problem: the problem of identifying trustworthy measures of consciousness[132]. Solving this problem is crucial, for detailed and comprehensive ToCs are unlikely to be of much use unless we also have the capacity to verify their predictions. It is useful to distinguish two (closely related) versions of the measurement problem. The first concerns the detection of conscious contents. Here, the primary challenge is to identify ways of distinguishing conscious from unconscious mental states that do not make controversial assumptions about the functional profile of consciousness (such as that conscious contents must be reportable or otherwise available for high-level cognitive control)[114,133,134]. The other version of the measurement problem focuses not on contents but on creatures. The questions here include how we might determine the distribution of consciousness in the animal kingdom[135]; whether certain classes of cerebral organoids[136] or artificial intelligence systems[135–137] are conscious; when consciousness first emerges in ontogenesis[138]; and when consciousness is retained in the context of traumatic brain injury[139]. Here, too, the challenge is to develop ways of measuring consciousness that avoid controversial assumptions about its functional profile (BOX 3).

Of course, the above challenges are already being addressed, to varying extents, by consciousness researchers. These efforts are now complemented by initiatives such as the adversarial collaboration model, which is encouraging proponents of ToCs to devise experiments with the specific goal of differentiating between alternative ToCs[17]. Consciousness remains scientifically controversial, yet there is every reason to think that the iterative development, testing and comparison of ToCs will lead to a much deeper understanding of this most profound of mysteries.

Published online 3 May 2022

Computational (neuro) phenomenology
The use of computational models to account for the phenomenal character of a conscious state in terms of (neural) mechanisms.

The measurement problem
The problem of identifying whether a particular mental state is conscious, or determining whether an organism or other system is, or has the capacity to be, conscious.

Cerebral organoids
Laboratory-grown neural structures that self-organize into systems with cellular and network features resembling aspects of the developing human brain.

1. Crick, F. & Koch, C. Towards a neurobiological theory of consciousness. *Semin. Neurosci.* **2**, 263–275 (1990).
2. Metzinger, T. (ed.) *Neural Correlates of Consciousness: Empirical and Conceptual Questions* (MIT Press, 2000).
3. Koch, C., Massimini, M., Boly, M. & Tononi, G. Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* **17**, 307–321 (2016).
4. de Graaf, T. A., Hsieh, P. J. & Sack, A. T. The 'correlates' in neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36**, 191–197 (2012).
5. Aru, J., Bachmann, T., Singer, W. & Melloni, L. Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36**, 737–746 (2012).
6. Tsuchiya, N., Wilke, M., Frassle, S. & Lamme, V. A. No-report paradigms: extracting the true neural correlates of consciousness. *Trends Cogn. Sci.* **19**, 757–770 (2015).
7. Klein, C., Hohwy, J. & Bayne, T. Explanation in the science of consciousness: from the neural correlates of consciousness (NCCs) to the difference-makers of consciousness (DMCs). *Philos. Mind Sci*. https://doi.org/10.33735/phimisci.2020.II.60 (2020).
8. Michel, M. et al. Opportunities and challenges for a maturing science of consciousness. *Nat. Hum. Behav.* **3**, 104–107 (2019).
9. Seth, A. K. Consciousness: the last 50 years (and the next). *Brain Neurosci. Adv.* **2**, 2398212818816019 (2018).
10. Seth, A. K. Explanatory correlates of consciousness: theoretical and computational challenges. *Cogn. Comput.* **1**, 50–63 (2009).
11. Searle, J. *The Rediscovery of the Mind* (MIT Press, 1992).
12. Varela, F. J. Neurophenomenology: a methodological remedy for the hard problem. *J. Conscious. Stud.* **3**, 330–350 (1996).
13. Seth, A. K. *Being You: A New Science of Consciousness* (Faber & Faber, 2021).
14. Dennett, D. C. Welcome to strong illusionism. *J. Conscious. Stud.* **26**, 48–58 (2019).
15. Frankish, K. *Illusionism as a Theory of Consciousness* (Imprint Academic, 2017).
16. Wiese, W. The science of consciousness does not need another theory, it needs a minimal unifying model. *Neurosci. Conscious.* **2020**, niaa013 (2020).
17. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).
This work sets out how an adversarial collaboration is planning to arbitrate between integrated information and global workspace ToCs.
18. Hameroff, S. & Penrose, R. Consciousness in the universe: a review of the 'Orch OR' theory. *Phys. Life Rev.* **11**, 39–78 (2014).
19. Chalmers, D. J. & McQueen, K. in *Quantum Mechanics and Consciousness* (ed Gao, S.) (Oxford Univ. Press, 2022).
20. Nagel, T. What is it like to be a bat? *Philos. Rev.* **83**, 435–450 (1974).
21. Bayne, T., Hohwy, J. & Owen, A. M. Are there levels of consciousness? *Trends Cogn. Sci.* **20**, 405–413 (2016).
This work challenges the common unidimensional notion of 'level of consciousness', outlining an alternative, richer, multidimensional account.
22. Metzinger, T. *Being No-One* (MIT Press, 2003).
23. Damasio, A. *Self Comes To Mind: Constructing the Conscious Brain* (William Heinemann, 2010).
24. Park, H. D. & Tallon-Baudry, C. The neural subjective frame: from bodily signals to perceptual consciousness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130208 (2014).
25. Bayne, T. *The Unity of Consciousness* (Oxford Univ. Press, 2010).
26. Bayne, T. & Chalmers, D. J. in *The Unity of Consciousness: Binding, Integration, and Dissociation* (ed Cleeremans, A.) 23–58 (Oxford Univ. Press, 2003).
27. Cummins, R. Functional analysis. *J. Philos.* **72**, 741–765 (1975).
28. Blake, R., Brascamp, J. & Heeger, D. J. Can binocular rivalry reveal neural correlates of consciousness? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130211 (2014).
29. Signorelli, C. M., Szczotka, J. & Prentner, R. Explanatory profiles of models of consciousness — towards a systematic classification. *Neurosci. Conscious.* **2021**, niab021 (2021).
30. Lau, H. & Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* **15**, 365–373 (2011).
This work presents a summary of empirical evidence favouring higher-order ToCs.
31. Rosenthal, D. *Consciousness and Mind* (Clarendon, 2005).
32. Brown, R. The HOROR theory of phenomenal consciousness. *Philos. Stud.* **172**, 1783–1794 (2015).
33. Cleeremans, A. Consciousness: the radical plasticity thesis. *Prog. Brain Res.* **168**, 19–33 (2008).
34. Cleeremans, A. et al. Learning to be conscious. *Trends Cogn. Sci.* **24**, 112–123 (2020).
35. Fleming, S. M. Awareness as inference in a higher-order state space. *Neurosci. Conscious.* **2020**, niz020 (2020).
36. Lau, H. Consciousness, metacognition, and perceptual reality monitoring. Preprint at *ArXiv* https://doi.org/10.31234/osf.io/ckbyf (2020).
37. Gershman, S. J. The generative adversarial brain. *Front. Artif. Intell.* https://doi.org/10.3389/frai.2019.00018 (2019).
38. Cohen, M. A., Dennett, D. C. & Kanwisher, N. What is the bandwidth of perceptual experience? *Trends Cogn. Sci.* **20**, 324–335 (2016).
39. Haun, A. M., Tononi, G., Koch, C. & Tsuchiya, N. Are we underestimating the richness of visual experiences? *Neurosci. Conscious.* **3**, 1–4 (2017).
40. Odegaard, B., Chang, M. Y., Lau, H. & Cheung, S. H. Inflation versus filling-in: why we feel we see more than we actually do in peripheral vision. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* https://doi.org/10.1098/rstb.2017.0345 (2018).
41. LeDoux, J. E. & Brown, R. A higher-order theory of emotional consciousness. *Proc. Natl Acad. Sci. USA* **114**, E2016–E2025 (2017).
42. Morrison, J. Perceptual confidence. *Anal. Philos.* **78**, 99–147 (2016).
43. Peters, M. A. K. Towards characterizing the canonical computations generating phenomenal experience. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/bqfr6 (2021).
44. Rosenthal, D. Consciousness and its function. *Neuropsychologia* **46**, 829–840 (2008).
45. Charles, L., Van Opstal, F., Marti, S. & Dehaene, S. Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage* **73**, 80–94 (2013).
46. Brown, R., Lau, H. & LeDoux, J. E. Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* **23**, 754–768 (2019).
47. Baars, B. J. *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1988).
48. Dehaene, S. & Changeux, J. P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227 (2011).
49. Mashour, G. A., Roelfsema, P., Changeux, J. P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).
This work presents a summary of the neuronal GWT and its supporting evidence.
50. Dehaene, S., Sergent, C. & Changeux, J. P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc. Natl Acad. Sci. USA* **100**, 8520–8525 (2003).
51. Naccache, L. Why and how access consciousness can account for phenomenal consciousness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* https://doi.org/10.1098/rstb.2017.0357 (2018).
52. Mashour, G. A. Cognitive unbinding: a neuroscientific paradigm of general anesthesia and related states of unconsciousness. *Neurosci. Biobehav. Rev.* **37**, 2751–2759 (2013).
53. Demertzi, A. et al. Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Sci. Adv.* **5**, eaat7603 (2019).
This large empirical study of functional connectivity patterns across different global states of consciousness focuses on how these patterns relate to underlying structural connectivity.
54. Barttfeld, P. et al. Signature of consciousness in the dynamics of resting-state brain activity. *Proc. Natl Acad. Sci. USA* **112**, 887–892 (2015).
55. Uhrig, L. et al. Resting-state dynamics as a cortical signature of anesthesia in monkeys. *Anesthesiology* **129**, 942–958 (2018).
56. Carruthers, P. *Human and Animal Minds: The Consciousness Questions Laid to Rest* (Oxford Univ. Press, 2019).
57. Tononi, G. Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* **215**, 216–242 (2008).
58. Tononi, G. Integrated information theory of consciousness: an updated account. *Arch. Ital. Biol.* **150**, 293–329 (2012).
59. Tononi, G., Boly, M., Massimini, M. & Koch, C. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* **17**, 450–461 (2016).
This work presents an account of the core claims and concepts of the integrated information ToC.
60. 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).
61. Tononi, G. & Koch, C. Consciousness: here, there and everywhere? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* https://doi.org/10.1098/rstb.2014.0167 (2015).
62. Haun, A. M. & Tononi, G. Why does space feel the way it does? Towards a principled account of spatial experienc. *Entropy* **21**, 1160 (2019).
63. 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).
64. Marshall, W., Gomez-Ramirez, J. & Tononi, G. Integrated information and state differentiation. *Front. Psychol.* **7**, 926 (2016).
65. Lamme, V. A. Towards a true neural stance on consciousness. *Trends Cogn. Sci.* **10**, 494–501 (2006).
66. Lamme, V. A. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
67. Hohwy, J. & Seth, A. K. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* **1**, 3 (2020).
68. Lamme, V. A., Super, H., Landman, R., Roelfsema, P. R. & Spekreijse, H. The role of primary visual cortex (V1) in visual awareness. *Vis. Res.* **40**, 1507–1521 (2000).
69. Pascual-Leone, A. & Walsh, V. Fast backprojections from the motion to the primary visual area necessary for visual awareness. *Science* **292**, 510–512 (2001).
This early study uses transcranial magnetic stimulation to reveal a role for re-entrant activity in conscious visual perception in humans.
70. Boehler, C. N., Schoenfeld, M. A., Heinze, H. J. & Hopf, J. M. Rapid recurrent processing gates awareness in primary visual cortex. *Proc. Natl Acad. Sci. USA* **105**, 8742–8747 (2008).
71. Lamme, V. A. How neuroscience will change our view on consciousness. *Cogn. Neurosci.* **1**, 204–220 (2010).
72. von Helmholtz, H. *Handbuch der Phsyiologischen Optik* [German] (Voss, 1867).
73. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
This work presents a classic exposition of predictive processing and its relevance for perception, cognition and action.
74. Friston, K. J. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
75. Seth, A. K. in *Open MIND* (eds Windt, J. M. & Metzinger, T.) (MIND Group, 2015).
76. Friston, K. J. Am I self-conscious? (Or does self-organization entail self-consciousness?). *Front. Psychol.* **9**, 579 (2018).
77. Seth, A. K. & Tsakiris, M. Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* **22**, 969–981 (2018).
78. Bruineberg, J., Dolega, K., Dewhurst, J. & Baltieri, M. The Emperor's new Markov blankets. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X21002351 (2021).
79. Hohwy, J. *The Predictive Mind* (Oxford Univ. Press, 2013).
80. Rao, R. P. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
81. Teufel, C. & Fletcher, P. C. Forms of prediction in the nervous system. *Nat. Rev. Neurosci.* **21**, 231–242 (2020).
82. Friston, K. J., Daunizeau, J., Kilner, J. & Kiebel, S. J. Action and behavior: a free-energy formulation. *Biol. Cybern.* **102**, 227–260 (2010).
83. Parr, T. & Friston, K. J. Generalised free energy and active inference. *Biol. Cybern.* **113**, 495–513 (2019).
84. Pennartz, C. M. A. Consciousness, representation, action: the importance of being goal-directed. *Trends Cogn. Sci.* **22**, 137–153 (2018).
85. Williford, K., Bennequin, D., Friston, K. & Rudrauf, D. The projective consciousness model and phenomenal selfhood. *Front. Psychol.* **9**, 2571 (2018).

86. Hohwy, J. New directions in predictive processing. *Mind Lang.* **35**, 209–223 (2020).
87. Seth, A. K. A predictive processing theory of sensorimotor contingencies: explaining the puzzle of perceptual presence and its absence in synesthesia. *Cogn. Neurosci.* **5**, 97–118 (2014).
88. O'Regan, J. K. & Noë, A. A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* **24**, 939–973; discussion 973–1031 (2001).
This primary description of the sensorimotor ToC argues that conscious perception is intimately related to action.
89. Seth, A. K. Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* **17**, 565–573 (2013).
This work presents a theoretical application of predictive processing to interoception and physiological regulation, relating this to experiences of emotion and selfhood.
90. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1833 (2017).
91. Solms, M. The hard problem of consciousness and the free energy principle. *Front. Psychol.* **9**, 2714 (2018).
92. Hohwy, J., Roepstorff, A. & Friston, K. Predictive coding explains binocular rivalry: an epistemological review. *Cognition* **108**, 687–701 (2008).
93. Parr, T., Corcoran, A. W., Friston, K. J. & Hohwy, J. Perceptual awareness and active inference. *Neurosci. Conscious.* **2019**, niz012 (2019).
94. Friston, K. J., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active inference: a process theory. *Neural Comput.* **29**, 1–49 (2017).
95. Boly, M. et al. Preserved feedforward but impaired top-down processes in the vegetative state. *Science* **332**, 858–862 (2011).
This neuroimaging study uses dynamic causal modelling to show that loss of consciousness in the vegetative state is associated with impaired top-down connectivity from frontal to temporal cortices.
96. Parr, T. & Friston, K. J. Working memory, attention, and salience in active inference. *Sci. Rep.* **7**, 14678 (2017).
97. Chalmers, A. *What is This Thing Called Science?* (Queensland Univ. Press, 2013).
98. Godfrey-Smith, P. G. *Theory and Reality: An Introduction to the Philosophy of Science* 2nd edn (Univ. Chicago Press, 2021).
99. Lipton, P. *Inference to the Best Explanation* (Routledge, 2004).
100. Lau, H. & Passingham, R. E. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc. Natl Acad. Sci. USA* **103**, 18763–18768 (2006).
This empirical study compares conscious and unconscious visual perception in humans, controlling for performance, and reveals differences in prefrontal activation.
101. van Vugt, B. et al. The threshold for conscious report: signal loss and response bias in visual and frontal cortex. *Science* **360**, 537–542 (2018).
This empirical study tracks the time course of neural signals in primate frontal cortex, showing that perceived stimuli elicit sustained activity, when compared with non-perceived stimuli.
102. Gaillard, R. et al. Converging intracranial markers of conscious access. *PLoS Biol.* **7**, e61 (2009).
103. Panagiotaropoulos, T. I., Deco, G., Kapoor, V. & Logothetis, N. K. Neuronal discharges and gamma oscillations explicitly reflect visual consciousness in the lateral prefrontal cortex. *Neuron* **74**, 924–935 (2012).
104. Kapoor, V. et al. Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat. Comm.* **13**, 1535 (2022).
105. Bellet, J. et al. Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neurosci. Conscious.* **2022**, niac005 (2022).
106. Levinson, M., Podvalny, E., Baete, S. H. & He, B. J. Cortical and subcortical signatures of conscious object recognition. *Nat. Commun.* **12**, 2930 (2021).
107. Boly, M. et al. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence. *J. Neurosci.* **37**, 9603–9613 (2017).
108. Raccah, O., Block, N. & Fox, K. C. R. Does the prefrontal cortex play an essential role in consciousness? Insights from intracranial electrical stimulation of the human brain. *J. Neurosci.* **41**, 2076–2087 (2021).
109. Odegaard, B., Knight, R. T. & Lau, H. Should a few null findings falsify prefrontal theories of conscious perception? *J. Neurosci.* **37**, 9593–9602 (2017).
110. Brascamp, J., Blake, R. & Knapen, T. Negligible fronto-parietal BOLD activity accompanying unreportable switches in bistable perception. *Nat. Neurosci.* **18**, 1672–1678 (2015).
This empirical 'no-report' study shows that fronto-parietal activity does not track switches in perceptual dominance when subjective reports are not required.
111. Sergent, C. et al. Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nat. Commun.* **12**, 1149 (2021).
112. Siclari, F. et al. The neural correlates of dreaming. *Nat. Neurosci.* **20**, 872–878 (2017).
113. Wong, W. et al. The Dream Catcher experiment: blinded analyses failed to detect markers of dreaming consciousness in EEG spectral power. *Neurosci. Conscious.* **2020**, niaa006 (2020).
114. Block, N. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behav. Brain Sci.* **30**, 481–548 (2007).
This work argues that research in psychology and neuroscience shows that there is a real and not merely conceptual distinction between phenomenal consciousness (that is, experience) and cognitive access to phenomenal consciousness.
115. Musgrave, A. in *Relativism and Realism in Science* (ed Nola, R.) 229–252 (Kluwer, 1988).
116. Song, C., Haun, A. M. & Tononi, G. Plasticity in the structure of visual space. *eNeuro* https://doi.org/10.1523/ENEURO.0080-17.2017 (2017).
117. Marshel, J. H. et al. Cortical layer-specific critical dynamics triggering perception. *Science* https://doi.org/10.1126/science.aaw5202 (2019).
118. Dembski, C., Koch, C. & Pitts, M. Perceptual awareness negativity: a physiological correlate of sensory consciousness. *Trends Cogn. Sci.* **25**, 660–670 (2021).
119. Sanchez, G., Hartmann, T., Fusca, M., Demarchi, G. & Weisz, N. Decoding across sensory modalities reveals common supramodal signatures of conscious perception. *Proc. Natl Acad. Sci. USA* **117**, 7437–7446 (2020).
120. Sergent, C. The offline stream of conscious representations. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* https://doi.org/10.1098/rstb.2017.0349 (2018).
121. Michel, M. & Doerig, A. A new empirical challenge for local theories of consciousness. *Mind Lang.* https://doi.org/10.1111/mila.12319 (2021).
122. Sergent, C. et al. Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Curr. Biol.* **23**, 150–155 (2013).
This empirical study reveals that conscious perception of a stimulus can be influenced by events happening (hundreds of milliseconds) after the stimulus appeared ('retro-perception').
123. Roseboom, W. et al. Activity in perceptual classification networks as a basis for human subjective time perception. *Nat. Commun.* **10**, 267 (2019).
124. Kent, L. & Wittmann, M. Special Issue: Consciousness science and its theories. Time consciousness: the missing link in theories of consciousness. *Neurosci. Conscious.* **2021**, niab011 (2021).
125. Husserl, E. *Ideas: A General Introduction to Pure Phenomenology* (Collier Books, 1963).
126. Yaron, I., Melloni, L., Pitts, M. & Mudrik, L. The ConTraSt database for analyzing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-021-01284-5 (2022).
This work presents an online resource of empirical studies of consciousness, organized with respect to different ToCs.
127. Joglekar, M. R., Mejias, J. F., Yang, G. R. & Wang, X. J. Inter-areal balanced amplification enhances signal propagation in a large-scale circuit model of the primate cortex. *Neuron* **98**, 222–234.e8 (2018).
128. VanRullen, R. & Kanai, R. Deep learning and the global workspace theory. *Trends Neurosci.* **44**, 692–704 (2021).
129. Shea, N. & Frith, C. D. The global workspace needs metacognition. *Trends Cogn. Sci.* **23**, 560–571 (2019).
130. Suzuki, K., Roseboom, W., Schwartzman, D. J. & Seth, A. K. A deep-dream virtual reality platform for studying altered perceptual phenomenology. *Sci. Rep.* **7**, 15982 (2017).
131. Vilas, M. G., Auksztulewicz, R. & Melloni, L. Active inference as a computational framework for consciousness. *Rev. Philos. Psychol.* https://doi.org/10.1007/s13164-021-00579-w (2021).
132. Browning, H. & Veit, W. The measurement problem in consciousness. *Philos. Top.* **48**, 85–108 (2020).
133. Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M. & Pessoa, L. Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends Cogn. Sci.* **12**, 314–321 (2008).
134. Michel, M. Calibration in consciousness science. *Erkenntnis* https://doi.org/10.1007/s10670-021-00383-z (2021).
135. Birch, J., Schnell, A. K. & Clayton, N. S. Dimensions of animal consciousness. *Trends Cogn. Sci.* **24**, 789–801 (2020).
136. Bayne, T., Seth, A. K. & Massimini, M. Are there islands of awareness? *Trends Neurosci.* **43**, 6–16 (2020).
This work presents an examination of the possibility of consciousness in isolated neural systems such as brain organoids, disconnected cortical hemispheres and ex cranio brains.
137. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
138. Hu, H., Cusack, R. & Naci, L. Typical and disrupted brain circuitry for conscious awareness in full-term and pre-term infants. (2021).
139. Owen, A. M. & Coleman, M. R. Detecting awareness in the vegetative state. *Ann. N Y Acad. Sci.* **9**, 130–138 (2008).
140. Cleeremans, A. The radical plasticity thesis: how the brain learns to be conscious. *Front. Psychol.* **2**, 86 (2011).
141. Jackendoff, R. *Consciousness and the Computational Mind* (MIT Press, 1987).
142. Prinz, J. *The Conscious Brain: How Attention Engenders Experience* (Oxford Univ. Press, 2012).
143. Chang, A. Y. C., Biehl, M., Yu, Y. & Kanai, R. Information closure theory of consciousness. *Front. Psychol.* **11**, 1504 (2020).
144. Tononi, G. & Edelman, G. M. Consciousness and complexity. *Science* **282**, 1846–1851 (1998).
This work presents an early proposal of how measures of neural complexity might relate to phenomenological properties of (all) conscious experiences.
145. Edelman, G. M. *Neural Darwinism: The Theory of Neuronal Group Selection* (Basic Books 1987).
146. Edelman, G. M. *The Remembered Present* (Basic Books, 1989).
147. Damasio, A. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness* (Harvest Books, 2000).
148. Graziano, M. S. A. The attention schema theory: a foundation for engineering artificial consciousness. *Front. Robot. AI* **4**, 60 (2017).
149. Dennett, D. C. *Consciousness Explained* (Little, Brown, 1991).
150. Ginsburg, S. & Jablonka, E. *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness* (MIT Press, 2019).
151. Aru, J., Suzuki, M. & Larkum, M. E. Cellular mechanisms of conscious processing. *Trends Cogn. Sci.* **24**, 814–825 (2020).
152. McFadden, J. Integrating information in the brain's EM field: the cemi field theory of consciousness. *Neurosci. Conscious.* **2020**, niaa016 (2020).
153. Fleming, S. M., Ryu, J., Golfinos, J. G. & Blackmon, K. E. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain* **137**, 2811–2822 (2014).
154. Fox, K. C. R. et al. Intrinsic network architecture predicts the effects elicited by intracranial electrical stimulation of the human brain. *Nat. Hum. Behav.* **4**, 1039–1052 (2020).
155. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).
156. Sergent, C., Baillet, S. & Dehaene, S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat. Neurosci.* **8**, 1391–1400 (2005).
157. Mediano, P. A. M., Seth, A. K. & Barrett, A. B. Measuring integrated information: comparison of candidate measures in theory and simulation. *Entropy* **21**, 17 (2019).
158. Casali, A. G. et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* **5**, 198ra105 (2013).
This empirical study shows that a measure of the complexity of the cortical response to transcranial magnetic stimulation distinguishes between a range of global conscious states, including disorders of consciousness.

159. Luppi, A. I. et al. Consciousness-specific dynamic interactions of brain integration and functional diversity. *Nat. Commun.* **10**, 4616 (2019).
160. Hardstone, R. et al. Long-term priors influence visual perception through recruitment of long-range feedback. *Nat. Commun.* **12**, 6288 (2021).
161. de Lange, F. P., Heilbron, M. & Kok, P. How do expectations shape perception? *Trends Cogn. Sci.* **22**, 764–779 (2018).
162. Melloni, L., Schwiedrzik, C. M., Muller, N., Rodriguez, E. & Singer, W. Expectations change the signatures and timing of electrophysiological correlates of perceptual awareness. *J. Neurosci.* **31**, 1386–1396 (2011).
This empirical study uses a perceptual hysteresis paradigm to show that expectations enhance and accelerate conscious perception.
163. Pinto, Y., van Gaal, S., de Lange, F. P., Lamme, V. A. & Seth, A. K. Expectations accelerate entry of visual stimuli into awareness. *J. Vis.* **15**, 13 (2015).
164. Chalmers, D. J. Facing up to the problem of consciousness. *J. Conscious. Stud.* **23**, 200–219 (1995).
This work presents the classic statement of the philosophical distinction between the 'hard' and 'easy' problems of consciousness.
165. Levine, J. Materialism and qualia: the explanatory gap. *Pac. Philos. Q.* **64**, 354–361 (1983).
166. Seth, A. K. *The Real Problem* (Aeon, 2016).
167. Balog, K. in *The Oxford Handbook of Philosophy of Mind* (eds Beckermann, A., McLaughlin, B. P., & Walter S.) 292–312 (Oxford Univ. Press, 2009).
168. Perry, J. *Knowledge, Possibility, and Consciousness* (MIT Press, 2001).
169. Varela, F. J., Thompson, E. & Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, 1993).
170. Carvalho, G. B. & Damasio, A. Interoception and the origin of feelings: a new synthesis. *Bioessays* **43**, e2000261 (2021).
171. Solms, M. *The Hidden Spring: A Journey to the Source of Consciousness* (Profile Books, 2021).
172. Merker, B. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* **30**, 63–81; discussion 81–134 (2007).
173. Parvizi, J. & Damasio, A. Consciousness and the brainstem. *Cognition* **79**, 135–160 (2001).
174. Naber, M., Frassle, S. & Einhauser, W. Perceptual rivalry: reflexes reveal the gradual nature of visual awareness. *PLoS ONE* **6**, e20910 (2011).
175. Casarotto, S. et al. Stratification of unresponsive patients by an independently validated index of brain complexity. *Ann. Neurol.* **80**, 718–729 (2016).
176. Shea, N. & Bayne, T. The vegetative state and the science of consciousness. *Br. J. Philos. Sci.* **61**, 459–484 (2010).
177. Birch, J. The search for invertebrate consciousness. *Noûs* **56**, 133–153 (2020).
178. Phillips, I. The methodological puzzle of phenomenal consciousness. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* https://doi.org/10.1098/rstb.2017.0347 (2018).

Acknowledgements
A.K.S. is Co-Director of, and T.B. is a Fellow in, the CIFAR Program on Brain, Mind, and Consciousness. A.K.S. is additionally grateful to the European Research Council (Advanced Investigator Grant 101019254) and the Dr. Mortimer and Theresa Sackler Foundation.

Author contributions
The authors both contributed to all aspects of preparing the article.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Neuroscience thanks L. Melloni, C. Sergent and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.