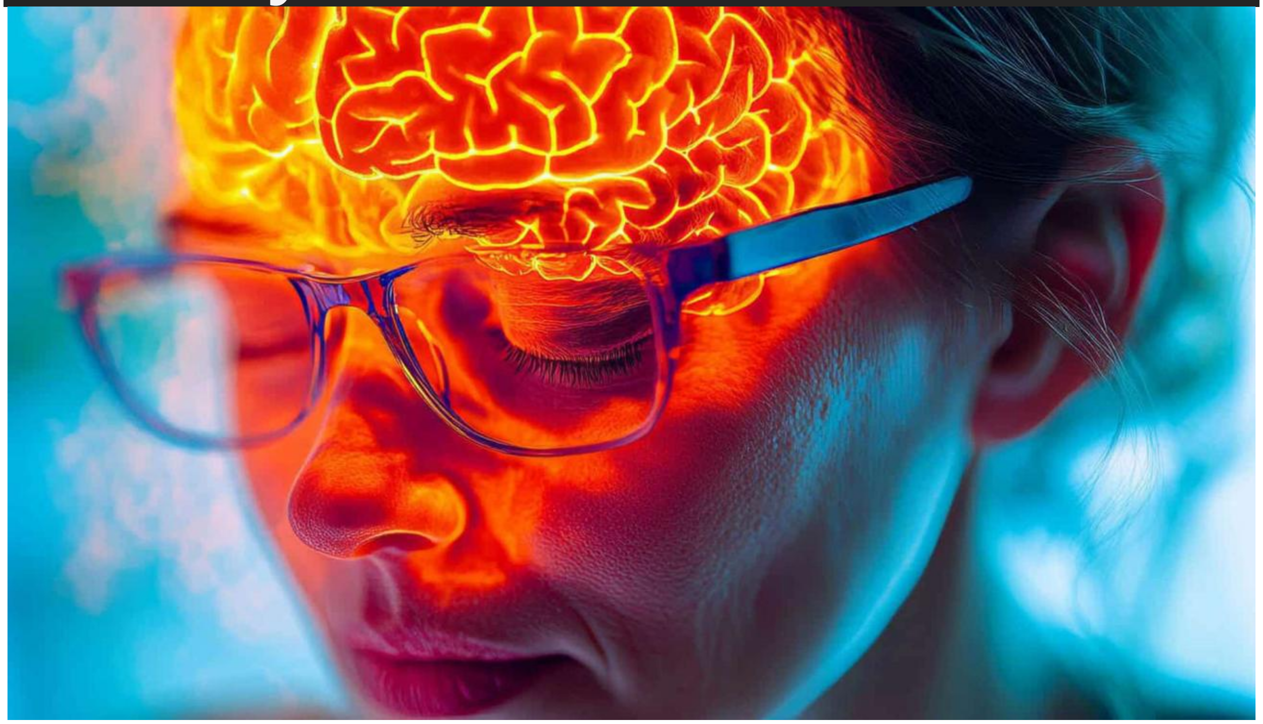
Consciousness explained: 12 new theories you haven't heard



A study published in the journal *Psychological Science* found that our brains perceive objects in everyday life of which we may never be aware.

Humans react, navigate, and decide with this invisible stream operating behind the scenes, a fact neuroscientists have known for decades.

Some of the theories we'll explore in this piece are decades old, some recent, but all reveal ways of understanding awareness that most of us have never encountered. From neural loops that echo perception to AI models simulating conscious agents, these twelve frameworks shed light on what it means to experience the world.

Global Workspace Theory

Studies show that roughly 50% of brief visual stimuli never reach conscious awareness, even when the eyes detect them.

Neuroscientist Stanislas Dehaene calls this phenomenon ignition, a sudden widespread activation in the fronto-parietal cortex that marks conscious access. Global Workspace Theory proposes that awareness arises when information is broadcast across multiple brain systems: perception, memory, and decision-making, enabling coordinated responses.

Stimuli that fail to ignite the workspace remain processed unconsciously, explaining why you might react to something without realizing it. **EEG and fMRI studies confirm that conscious perception correlates with measurable patterns of cortical activity.**

Bicameral Mentality   julianjaynes.org

Ancient texts, such as Homer's Iliad, describe warriors receiving direct instructions from gods.

Julian Jaynes hypothesized that these were not metaphorical but literal accounts of a bicameral mind, where decision-making occurred through auditory hallucinations rather than introspective thought.

Archaeological and linguistic evidence suggests that complex self-reflection emerged only around 3,000 years ago, coinciding with the development of written language.

Jaynes' theory frames consciousness as both a neurological and cultural evolution: humans learned to internalize commands and create a reflective mind. **While controversial, this perspective highlights that awareness is shaped by history and society, not only by neurons.**

Cultural Construction of Consciousness

Research in developmental and cross-cultural psychology shows that the way people talk about thoughts and feelings influences how they perceive their own minds. For example, children acquiring inner speech gradually develop self-reflective consciousness, while linguistic communities with fewer terms for introspection may report different experiences of awareness.

Julian Jaynes' theory aligns with this, suggesting that written language and cultural practices helped humans internalize thought rather than rely on the voices of gods.

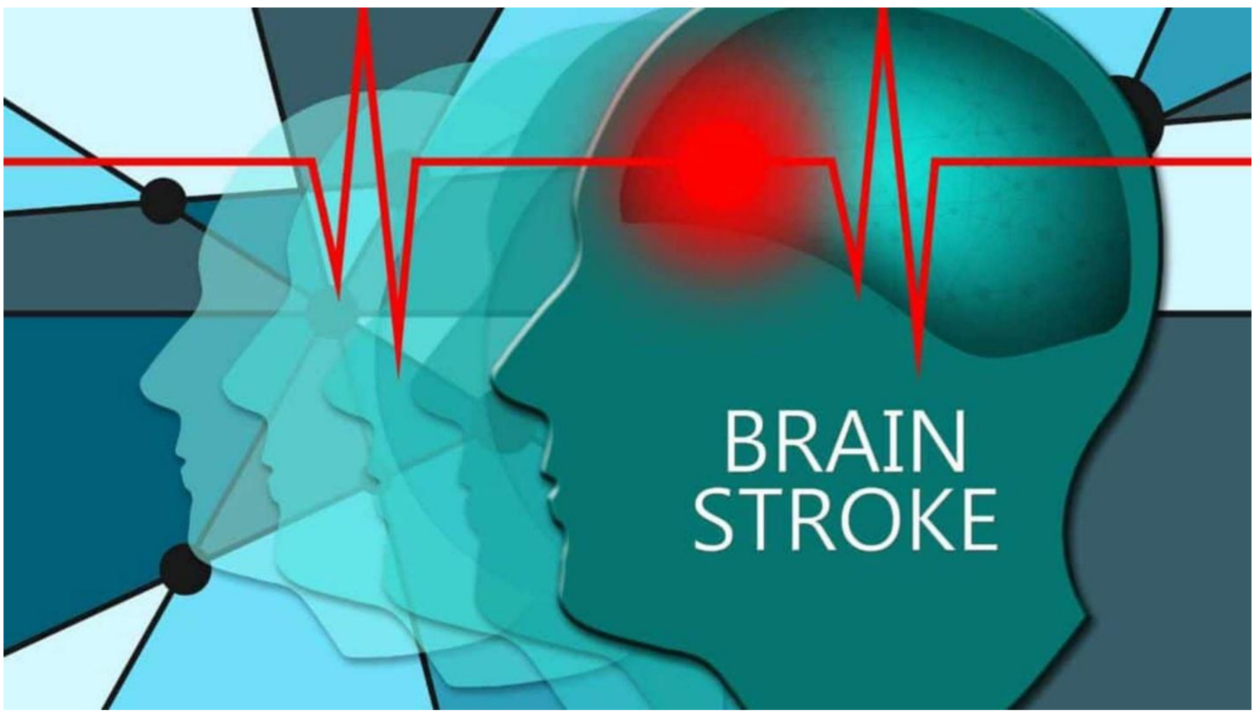
Modern studies confirm that meditation, literacy, and narrative training can measurably alter neural networks associated with self-awareness. Consciousness, therefore, is not only a biological phenomenon but also a cultural technology shaped by language, social interaction, and education.

Local Recurrence

When a visual stimulus hits the retina, signals bounce within local cortical circuits multiple times before reaching conscious awareness. **This repetition, called local recurrence, can last just tens of milliseconds, but it is crucial: experiments using**

backward masking show that stimuli disrupted before these loops fail to be consciously perceived.

Neuroscientist Victor Lamme proposes that consciousness emerges from these reverberating signals, rather than requiring global cortical broadcasting. In practice, this explains why you might sense movement or color without immediately recognizing it. Local recurrence demonstrates that awareness can originate in high-speed neural echo chambers, highlighting a subtle, rapid, and localized mechanism of perception.



Attention Schema Theory

Some stroke patients fail to notice objects on one side of their visual field, despite intact eyes, a condition called hemispatial neglect.

Attention Schema Theory, **proposed by Michael Graziano, explains this as the brain's internal model of its own attention.** The theory suggests consciousness arises when the brain represents where attention is directed, allowing flexible behavior and self-monitoring.

fMRI studies show that parietal and frontal networks track attentional focus, and damage to these networks impairs awareness. **Consciousness, under AST, is functional: it's the brain's way of keeping tabs on itself, turning perception into actionable insight.**

Conscious Agent Networks

Donald Hoffman proposes that perception is not a window onto reality but a user interface shaped by evolution. In simulations, conscious agents interact using simplified rules, and these interactions generate complex behaviors without any direct representation of objective reality.

Experimental AI models show that agents operating on fitness-optimized interfaces thrive even when their sensory inputs do not reflect true environmental states. Hoffman's theory reframes awareness as relational and strategic: your perceptions are icons on a survival dashboard, optimized for adaptive behavior rather than truth.

While counterintuitive, this computational perspective emphasizes that consciousness emerges from interaction networks rather than just brain structures, providing a testable framework that bridges psychology, evolution, and computer science.

Integrated Information Theory

Experiments using the **perturbational complexity index (PCI) show that patients in minimally conscious states often exhibit PCI values below 0.3, whereas fully awake individuals exhibit PCI values above 0.6.**

Integrated Information Theory (IIT), developed by Giulio Tononi, proposes that consciousness corresponds to the amount of integrated information, measured as phi, in a system.

Even simple neural networks could, theoretically, support minimal awareness if the causal structure is complex enough.

IIT transforms consciousness from a subjective experience into a measurable property of brain organization, bridging the gap between phenomenology and empirical neuroscience.

Integrated Information Theory

Giulio Tononi, the founder of Integrated Information Theory (IIT), **describes consciousness as a property of causal structure, not of matter itself.** IIT proposes that the degree of awareness corresponds to the amount of integrated information, measured mathematically as phi.

Clinical applications support this: **patients in vegetative states often have PCI (perturbational complexity index) values below 0.3, while healthy awake adults exceed 0.6**, showing a measurable correlation between brain integration and conscious experience.

IIT also predicts that even artificial networks could, in theory, possess minimal consciousness if their internal connectivity produces sufficient integration. This approach reframes consciousness from a purely subjective experience into a quantifiable property, providing a bridge between neuroscience, computation, and philosophy.

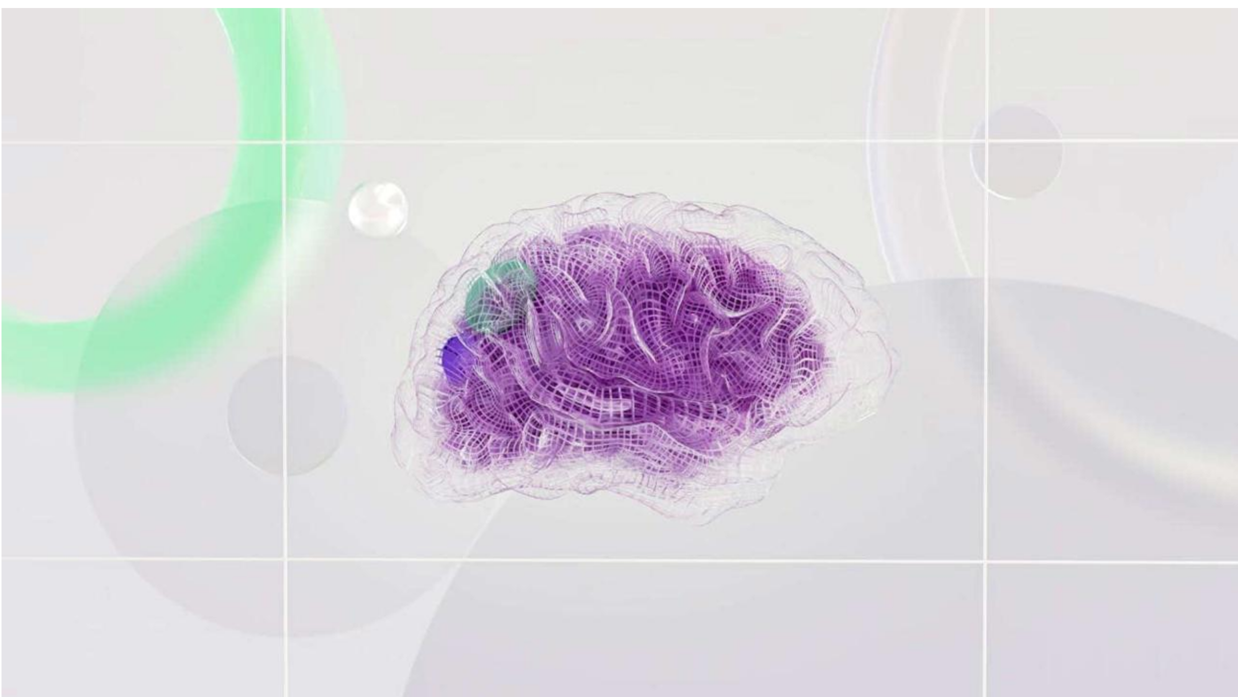
Critical Brain Hypothesis

Your brain operates near a delicate balance: too much order, and it becomes rigid; too much chaos, and signals

dissolve. Neuroscientists have found that neuronal networks at criticality, the boundary between order and disorder, maximize responsiveness and information processing.

Studies of brain activity in sleep, anesthesia, and coma show that consciousness diminishes when networks move away from this critical point. In these conditions, EEG readings reveal reduced neuronal avalanches, indicating less coordinated activity.

This suggests that **awareness emerges not from a single region but from the dynamic interplay of neurons poised at the edge of chaos**, explaining why minor disruptions can dramatically alter conscious experience.



Electromagnetic Field Theory

© Image Credit: Google DeepMind/pexels

When neurons fire, they create not only electrical signals but also overlapping electromagnetic (EM) fields. Researchers have observed that **altering these fields in experimental settings can shift perception**, **even when neural firing patterns remain unchanged.**

This suggests that EM fields may play a causal role in consciousness. For example, **studies using transcranial magnetic stimulation can transiently disrupt awareness in localized regions by manipulating the brain's EM patterns**, revealing the field's influence on experience.

Unlike theories that focus solely on neuron spikes or network structure, EM field theories propose that consciousness arises from the collective resonance of brain waves, offering a measurable physical substrate for awareness.

Orch-OR

© fizkes via Shutterstock.

Roger Penrose and Stuart Hameroff propose that consciousness arises from orchestrated quantum events in microtubules within neurons.

Unlike classical neural models, this theory links awareness to subcellular physical processes. Evidence remains limited, but studies suggest that microtubules may support coherent oscillations that can influence neural activity.

Orch-OR suggests that each moment of consciousness corresponds to an objective reduction of quantum states, producing discrete experiential events. While highly debated, this framework challenges the assumption that classical computation alone generates awareness, offering a bridge between neuroscience and fundamental physics.

If correct, consciousness would be a phenomenon embedded in the very fabric of the brain's quantum machinery, not just its electrical activity.

Panprotopsychism

Some contemporary theorists suggest that consciousness may be a fundamental property of structured systems, not just brains. Modern panprotopsychism proposes that **complex organization, rather than material substance, can support rudimentary awareness**.

Neural networks, and even sufficiently organized artificial systems, could, in theory, exhibit proto-conscious experience if their causal structures are sufficiently complex. This approach reframes consciousness as an emergent property of interactions within a system, measurable and testable in principle, rather than an epiphenomenon of biology alone.

Philosophically, it challenges the traditional duality of mind and matter, inviting reflection on whether awareness is woven into the very architecture of the universe, rather than being confined to humans.

Key Takeaways

- **Consciousness is multi-dimensional:** Awareness cannot be fully explained by a single model; neuroscience, psychology, quantum physics, computation, and cultural evolution all offer valuable insights.
- **Awareness is measurable and observable:** Experiments with EEG, fMRI, TMS, PCI, and AI simulations show that consciousness can be linked to quantifiable neural patterns or functional models.
- **Function and perception are intertwined:** The brain's attentional systems, feedback loops, and evolved interfaces highlight that experience is often about **adaptive processing rather than the direct representation of reality**.
- **Consciousness spans biology, culture, and computation:** From bicameral minds and language shaping self-reflection to quantum microtubules and AI simulations, these theories illustrate that awareness emerges from interactions across multiple levels, not a single brain mechanism.

Neuroscience & Biobehavioral Reviews
Date: March 2025
Article: 106053
Volume: Volume 170
Published by

Review article
Open access

Unpacking the complexities of consciousness: Theories and reflections

Liad Mudrik [a b c], Melanie Boly [d], Stanislas Dehaene [c e f], Stephen M. Fleming [c g h i], Victor Lamme [j], Anil Seth [c k], Lucia Melloni [c l]
Show more

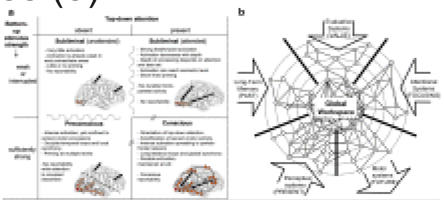
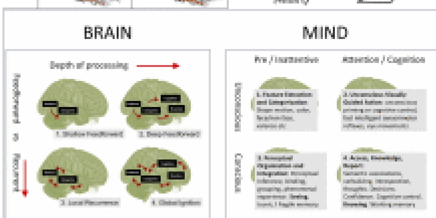
10.1016/j.neubiorev.2025.106053

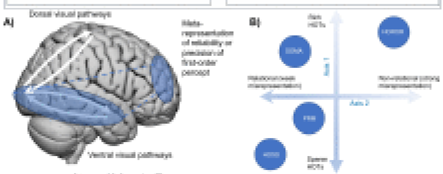
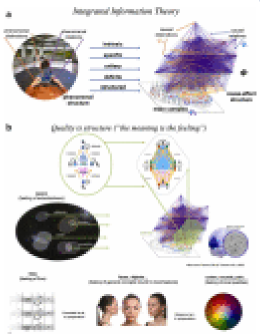
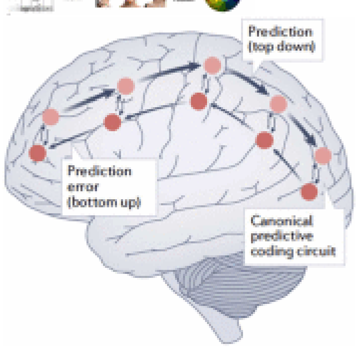
- **Article**
- Recommended articles
- Cited by (40)
- Metrics

Outline

Show full outline

Figures (5)

1. 
2. 

3. 
4. 
5. 

Tables (1)

1. Table 1. Key commitments of each theory outlining their core focus, metaphysical commitments, and explanations of non-conscious processes and key neural mechanisms. It also highlights the role of attention, co...

Highlights

- •

  We summarize a public debate between proponents of five theories of consciousness,
- •

  The theories are compared with respect to different key questions,
- •

  These include their explanandum, metaphysics, key mechanisms and testability,
- •

  Currently, there is more disagreement than agreement between theories,
- •

  This pertains to fundamental issues, including the definition of consciousness.

Abstract

As the field of consciousness science matures, the research agenda has expanded from an initial focus on the neural correlates of consciousness, to developing and testing theories of consciousness. Several theories have been put forward, each aiming to elucidate the relationship between consciousness and brain function. However, there is an ongoing, intense debate regarding whether these theories examine the same phenomenon. And, despite ongoing research efforts, it seems like the field has so far failed to converge around any single theory, and instead exhibits significant polarization. To advance this discussion, proponents of five prominent theories of consciousness—Global Neuronal Workspace Theory (GNWT), Higher-Order Theories (HOT), Integrated Information Theory (IIT), Recurrent Processing Theory (RPT), and Predictive Processing (PP)—engaged in a public debate in 2022, as part of the annual meeting of the Association for the Scientific Study of Consciousness (ASSC). They were invited to clarify the explananda of their theories, articulate the core mechanisms underpinning the corresponding explanations, and outline their foundational premises. This was followed by an open discussion that delved into the testability of these theories, potential evidence that could refute them, and areas of consensus and disagreement. Most importantly, the debate demonstrated that at this stage, there is more controversy than agreement between the theories, pertaining to the

most basic questions of what consciousness is, how to identify conscious states, and what is required from any theory of consciousness. Addressing these core questions is crucial for advancing the field towards a deeper understanding and comparison of competing theories.

Keywords

Theories of consciousness

;

Global neuronal workspace Theory

;

Integrated information theory recurrent processing theory

;

Higher order thought theories

;

Predictive processing theory

- Previous article in this issue
- Next article in this issue

1. Introduction

In recent years, as the field of consciousness studies matures, the research program is gradually transitioning from its initial focus on the search for the neural correlates of consciousness (Crick and Koch, 2003), and the accumulation of empirical observations, to developing theories of consciousness (ToCs). A shared research program of these ToCs is to provide an explicit account of the physical and psychological mechanisms of conscious experience (e.g., Doerig et al., 2020; Kuhn, 2024; Sattin et al., 2021; Seth and Bayne, 2022; Storm et al., 2024; Yaron et al., 2022). That is, these theories all strive to explain the mechanisms that underlie conscious, phenomenal experiences: many of them accordingly target what differentiates between processes that are accompanied by a conscious experience and those that are not, and some also ask what differentiates between two conscious experiences, though this question has been by and large less thoroughly dealt with thus far (Seth and Bayne, 2022). Yet despite substantial empirical and theoretical progress, there is still no clear path towards a unified account (Lepauvre and Melloni, 2021, Seth, 2018, Seth and Bayne, 2022). This failure to establish an agreed-upon account might stem from at least two reasons. First, the explanandum of the theories (i.e., the phenomena they aim at explaining) might differ (Sattin et al., 2021, Signorelli et al., 2021). If so, seeking a unified theory is likely to fail, and a more promising strategy would be to first delineate the different explananda and establish the boundary conditions for these theories, as a step towards developing a more integrative account (Evers et al., 2024). Second, the methodologies and measures used in the empirical work around the theories has been shown to differ, potentially leading to a confirmatory bias in the literature. Indeed, a recent study showed that support for ToCs can be predicted from the methodological choices of the experiments, irrespective of the obtained result (Yaron et al., 2022; see also Promet and Bachkann, 2022).

In response to this fragmented landscape, the Association for the Scientific Study of Consciousness (ASSC) in 2022 hosted the "Great Consciousness Debate," featuring proponents of four theories: Stanislas Dehaene presented the Global Neuronal Workspace Theory (GNWT; Dehaene and Naccache, 2001; Mashour et al., 2020), Melanie Boly presented the Integrated Information Theory (IIT; Albantakis et al., 2023; Tononi, 2008), Stephen Fleming presented Higher Order theories (HOTs; Brown et al., 2019; Lau and Rosenthal, 2011), and Victor Lamme presented the Recurrent Processing Theory (RPT; Lamme, 2006; Lamme and Roelfsema, 2000). Anil Seth had planned to present the Predictive Processing Theory (PPT; Hohwy and Seth, 2020; Seth, 2021), but was unable to attend. Thus, his contribution to this paper was provided in retrospect. Importantly, as this paper is based on the 2022 debate, it only features the five theories invited to participate in that debate. These theories only subtend some of the theoretical landscape in the field, which includes other theories (for reviews, see again Doerig et al., 2020, Kuhn, 2024, Sattin et al., 2021, Seth and Bayne, 2022, Storm et al., 2024). For example, all the theories discussed here are primarily corticocentric, as they rely mainly on cortical mechanisms (and some also on underlying thalamocortical loops), unlike other theories that focus on subcortical substrates (e.g., Merker, 2007; Ward, 2011; see also Solms, 2019). Accordingly, we do not consider them to fully represent the breadth of theories and views, and we welcome similar future debates amongst other prominent theories.

The debate aimed at updating the academic community on each theory, identifying commonalities that might pave the way towards a unified theory, and highlighting conceptual differences to guide empirical testing. Each proponent was asked to define their theory's explananda, its core mechanisms and foundational premises, as well as to suggest results which, if established, could potentially alter their stance on their theory. In this paper, we first provide summaries of each proponent's presentation. These summaries are largely based on the talks given during the debate, with few modifications introduced during the writing process by each of the proponents to improve readability. Table 1 further provides a concise summary prepared by the proponents, outlining for each theory its core focus, metaphysical commitments, and explanations of non-conscious processes and key neural mechanisms. It also highlights assumptions regarding the interplay between consciousness, attention, cognition, and evolutionary factors, as well as potential challenges for each theory. We then follow with a section written primarily by the mediators of the debate, highlighting what they considered key insights and challenges drawn from the open discussion between the proponents. There, we primarily focus on the ongoing effort to clarify concepts, detect points of disagreement as well as find common ground among the leading theories of consciousness (for other attempts to do so, see Chis-Ciure et al., 2024; Doerig et al., 2020; Signorelli et al., 2021; Storm et al., 2024).

Table 1. Key commitments of each theory outlining their core focus, metaphysical commitments, and explanations of non-conscious processes and key neural mechanisms. It also highlights the role of attention, cognition, and evolutionary factors within each theory, as well as potential challenges. This comparison offers a structured insight into the strengths and weaknesses of different theoretical approaches to consciousness.

Theory	Explanandum	What needs to be explained	Metaphysical commitment	Non-conscious information & associated mechanism	Key neural feature	Key comput featu
GNWT	Conscious processing (i.e., which processes require or do not require consciousness), and conscious	Any reportable perceptual or non-perceptual experience (e.g., experience of knowing that one just made an error).	Functionalism	Dehaene and Naccache (2001) and Dehaene et al., (2006)(Box 2) introduced a taxonomy of non-conscious information	Global ignition defined as non-linear activation of neurons in Prefrontal-Parietal and Cingulate	Flexible shar information a otherwise mo brain system

Theory	Explanandum	What needs to be explained	Metaphysical commitment	Non-conscious information & associated mechanism	Key neural feature	Key comput featur
	phenomenology (i.e., what a participant experiences at a given moment)	Any delay or change in this experience (e.g., attentional blink or psychological refractory period). Any contrast between reportable versus non-reportable aspects of perceptual and cognitive processing		in the brain, including (1) information not represented in cortical patterns (e.g. dormant synaptic connectivity); (2) information not represented by an explicit pattern of firing; (3) information encoded in regions disconnected from GNW areas; (4) activation evoked by weak sensory stimuli; (5) activation evoked by strong stimuli that failed to mobilize top-down attention and remain associated with feedforward processing or confined to local interactions within sensory regions, without triggering ignition	cortex, and broadcasting of those signals back to other distant areas	
RPT	Conscious phenomenology (i.e., what a participant experiences at a given moment)	Sensation (vision in particular), perceptual organization and unification or integration, perceptual inference, perceptual richness, cognitive impenetrability	Functionalism (with structural implementation: the to-be-explained functions happen to be mediated by recurrent interactions in most brains)	Feedforward processing sweep, extracting low (shape, color, motion etc.) and high-level (faces, objects, scene gist) features, and activating sensorimotor reflexes up to cognitive control functions	The distinction between Fast Feedforward Sweep (FFS) and Recurrent Processing (RP), the first unconscious, the latter yielding interactions between visual areas and hence conscious sensation. These interactions are for example expressed in feedback mediated	Perceptual organization integration, a inference by of interaction between units representing individual fea Possibly perc learning

Theory	Explanandum	What needs to be explained	Metaphysical commitment	Non-conscious information & associated mechanism	Key neural feature	Key comput featur
					modulations of early visual cortex (contextual modulation), effects of Transcranial Magnetic Stimulation (TMS) on early visual cortex, effects of masking, anesthesia, etc. showing a selective disruption of RP but not FFS. Possibly selective N-methyl-D-aspartate (NMDA) receptor activation by RP	
HOTs	Conscious phenomenology (i.e., presence vs. absence of phenomenal experience)	Dissociations between phenomenology and behavioral performance; relationships between phenomenology and perceptual metacognition; inner awareness of mental function; reality monitoring; subjective inflation	Functionalism	Any first-order representation without an accompanying higher-order representation	Neural substrates of subpersonal (implicit) metacognition within parietal and prefrontal cortex; for joint determination variants of HOT, joint contributions of first-order (e.g., visual) and higher-order areas to consciousness	HOT variants according to computationa features. The Higher-Order Space / Perce Reality Moni variants prop "lean" higher states, pointir the precision reliability of order representatio whereas the I Order Representatio Representatio variant posits content" high order represe

Theory	Explanandum	What needs to be explained	Metaphysical commitment	Non-conscious information & associated mechanism	Key neural feature	Key comput featur
IIT	What determines whether consciousness is present vs. absent; and what determines why specific experiences feel the way they do	First, what determines whether consciousness is present vs. absent? With respect to anatomy, why do certain part of the corticothalamic system seem to contribute directly to consciousness, while many other parts of the brain, such as the cerebellum and certain cortical areas do not? With respect to physiology, why is it that consciousness vanishes during deep sleep even though neurons continue to be active, and during generalized tonic-clonic seizures, even though neurons fire maximally and in a highly synchronous manner? Second, what determines why specific experiences feel the way they do? Why do visual space and body space feel extended, why does time feel flowing, and why do objects, colors, sounds or touch, feel the way they do?	The starting point of IIT is the existence of experience (0th axiom). This truth is not the result of an inference. It is immediate and irrefutable. The five axioms of IIT identify the essential properties of phenomenal existence—those that are true of every conceivable experience (every experience is intrinsic, specific, unitary, definite and structured). These essential phenomenal properties are then formulated in physical terms—in terms of a substrate we can observe and manipulate. For other assumptions, see Albantakis et al., (2023), and IIT WIKI (www.iit.wiki)	The contents of experience are fully specified by the shape of the cause-effect structure specified by the substrate of consciousness (a main complex with its particular spatiotemporal grain) in a particular state. Every other part of the brain (excluded from the main complex) is excluded and does not directly contribute to the contents of the experience specified by the system	Dense, divergent-convergent hierarchical 3D lattice of specialized units in a state of causal ‘readiness,’ regardless of activity	None. An exp is a cause-eff structure, not computation, function, or a process (Finc al., 2024; Ton al., in press; Zaeem and Tononi, 2
PP	Conscious phenomenology	Phenomenological character of	Materialism (subtypes may	Open question; potentially predictive	Neural activity corresponding	Hierarchical minimization

Theory	Explanandum	What needs to be explained	Metaphysical commitment	Non-conscious information & associated mechanism	Key neural feature	Key comput featur
	(i.e., presence and absence of consciousness and the specific phenomenal character of a given experience)	subjective experiences of world and self, their function, and their presence vs. absence. More specifically, how expectations shape or give rise to conscious contents, the functional role of top-down connectivity, what are the sufficient conditions for conscious perception in terms of predictive inference	have other commitments or suggest more specific metaphysical commitments, such as biological naturalism, computational functionalism, or non-computational functionalism)	inferences without an active component, or without a top-down generative component	to predictive inferences about the causes of sensory signals; potentially the active resolution of sensory prediction error through overt or covert action	precision wei prediction err more general active minim of variational expected free energy. Some versions of P not be compu per se, but m instead be dy theories (exp in terms of differential equations etc

2. Presentation of the debated theories

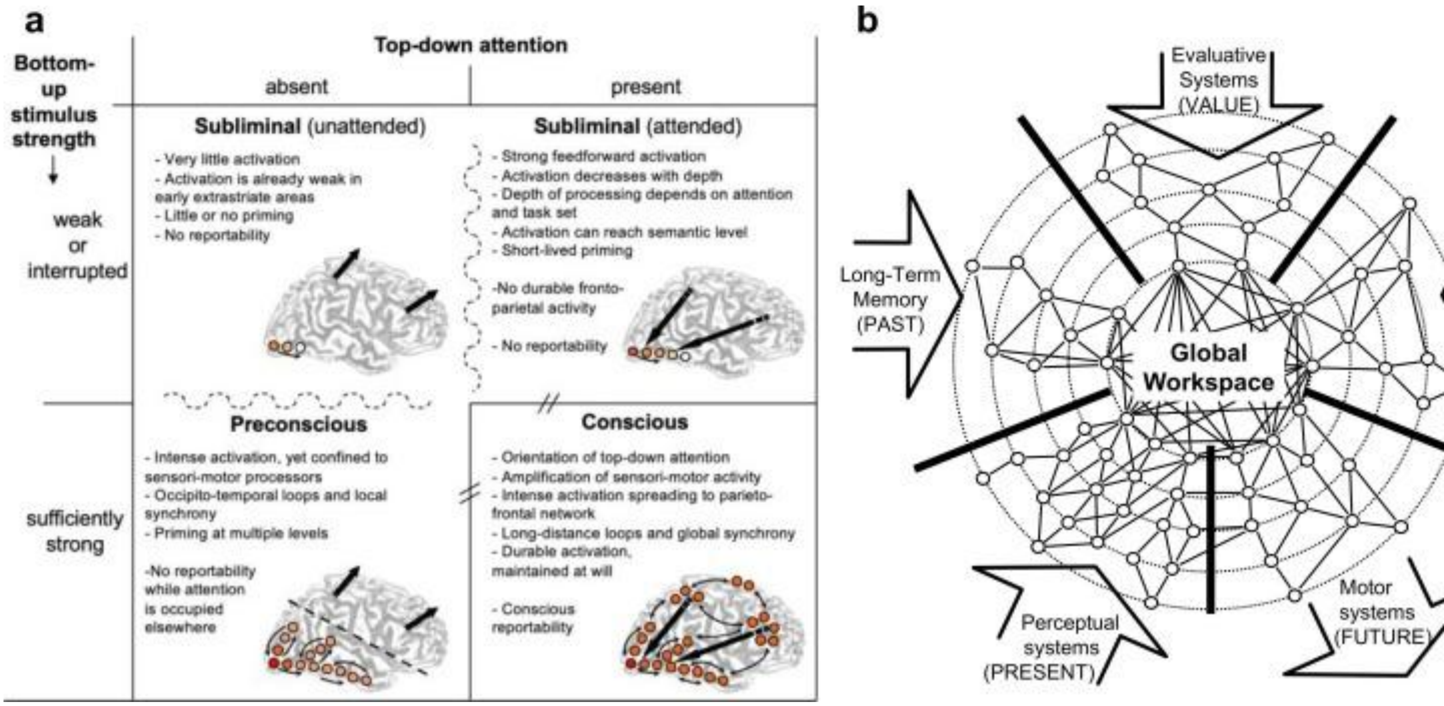
2.1. Global neuronal workspace / Stanislas Dehaene

The Global Neuronal Workspace theory (GNWT), which was proposed more than 25 years ago (Dehaene et al., 1998), holds that what we call consciousness is a computational property, characteristic of a certain type of information processing (Dehaene and Naccache, 2001). Tying the theory with current developments in Artificial Intelligence (AI) and Machine Learning (ML), the theory holds that in principle, the architecture underlying conscious processing could be implemented in machines. However, consciousness is a polysemic term. To clarify the situation, together with Sid Kouider and Hakwan Lau, we distinguished three different types of computations, with the latter two representing two different dimensions of consciousness (Dehaene et al., 2017): C0, which is characteristic of many existing algorithms, refers to *nonconscious computations* (i.e., information processing operations that can be carried out without consciousness). C1 represents the level of *conscious access* where a workspace globally broadcasts and amplifies a specific piece of information, akin to a routing system. C2 refers to the ability of a system to represent itself. A system with both self-representation and conscious access to it will have knowledge about itself: it knows what it knows, and it knows what it doesn't know. What is the scope of the GNW theory? The theory aims to account for two different things: One is *conscious processing* (i.e., which processes require or do not require consciousness), and the other is *conscious phenomenology* (i.e., what a participant experiences at a given moment). In the last 25 years, our understanding of conscious vs. nonconscious processing has expanded substantially. Many experiments have revealed that nonconscious processing is extensive, so that much can be done without awareness (Kouider and Dehaene, 2007, Mudrik and Deouell, 2022, Weiskrantz, 1997). Yet, we have also learned that something special happens when a piece of information becomes conscious. According to GNWT, this 'something special' is a non-linear ignition, leading to a global availability of information. There is a system of neurons that is able to select a piece of information, amplify it and broadcast it across the brain, and across modules. This broadcasting allows the sharing of information across otherwise nonconscious processing pathways, which enables the flexible recombination and routing of information, including the ability to send it to circuits for verbal report. Thus, reportability is just one marker for information being available in the workspace. Global availability is the key difference between conscious and unconscious processes. According to GNWT, even detailed visual content, say a fine visual Vernier pattern, if it is within the focus of attention, should be accessed by GNW neurons and broadcasted to the rest of the brain. This is what allows us to report even on very fine perceptual details, if and only if we access them consciously.

Another important clarification should be made here; the workspace neurons are distributed. GNW is not a localist theory of consciousness. A common error is to equate it with claims about prefrontal areas. Much to the contrary, the GNW is postulated to rely on a robust, redundant, highly distributed system of neurons with long-distance axons, which incorporates both parietal and prefrontal cortices. During ignition by an external sensory stimulus, a subset of those neurons become activated (a subspace of the high-dimensional set of all possible GNW states), thus encoding the details of the conscious contents, while the other neurons, reflecting what the stimulus is not, are inhibited. In this manner, GNW areas amplify, hold online, and propagate the information originating for processors in sensory areas, and these regions in turn receive top-down signals from fronto-parietal neurons. Therefore, a key claim is that any conscious contents should be represented by a distributed assembly of neurons, present in both prefrontal cortex and these related areas. This prediction is well-supported by anatomical findings of long-distance connections interconnecting those regions and forming a central cortical core also supported by thalamo-cortical and basal ganglia loops (Markov et al., 2013). What should be common to all conscious-processing states is a processing style, or a type of neural trajectory within those areas, with non-linear ignition leading to metastable activity in these high-level areas lasting for about 100–200ms (King et al., 2014, Schurger et al., 2015).

GNW theory makes precise empirical predictions (Dehaene and Changeux, 1997, Dehaene et al., 2006, Dehaene and Naccache, 2001), which are also supported by simplified simulations of thalamo-cortical activity (Dehaene and Changeux, 2005, Dehaene et al., 1998, Dehaene et al., 2003, Klatzmann et al., 2023), many of which have already been borne out by the data.

First, the theory predicts that nonconscious processing can run very deep, along specialized and automatized processing chains, which may be subcortical as well as cortical (Fig. 1). According to the theory, knowledge encoded within the nervous system can remain nonconscious for a variety of reasons, and a taxonomy of nonconscious states has been proposed (Dehaene et al., 2006). The theory predicts that information necessarily remains nonconscious and cannot be consciously accessed (1) if it is merely encoded in latent form, by matrices of synaptic weights; (2) if it is not explicitly condensed in the firing of small specialized groups of neurons; (3) if those neurons are functionally disconnected from GNW neurons (e.g., those located in the brainstem); (4) if processing is confined to a brief, unstable traveling wave of firing, insufficient strong to cause ignition; or (5) if processing occurs while top-down attention is focused on another stimulus or task set. The theory stresses that information becomes conscious only when its mental representation receives top-down amplification and gets expanded into a global parietofrontal metastable state.



1. Download: Download high-res image (584KB)
2. Download: Download full-size image

Fig. 1. (a) This classic figure adapted from Dehaene et al. (2006) shows the joint need for bottom-up stimulus strength (vertical axis) and top-down amplification (horizontal axis) in creating the conditions needed for access of a visual stimulus to the global neural workspace. While manu chains of processing can occur non-consciously, the theory stipulates that a distributed system of workspace neurons with long-distance axons must be activated for a sufficient duration (minimun 100-200 ms) for conscious appraisal to occur. As illustrated in the bottom right cell, what we experience as consciousness is the availability of information. (b) No longer confined to a narrow specialized circuit, information which is consciously accessed becomes broadcasted globally and flexibly to many other brain processors, thus allowing the information to be named, memorized, and acted upon in a deliberate, intentional manner (adapted from Dehaene et al., 1998).

Second, there should be a sudden global and non-linear ignition, similar to a first-order phase transition, particularly but not exclusively in prefrontal cortex, whenever a novel content gains access to consciousness. This ignition should only be seen when the information is new and leads to a refresh of information within this global workspace.

Third, a crucial point for the theory is that such conscious ignition can be temporally decoupled from the external world: the onset and duration of conscious experience is not determined by the external stimulus, but by the availability of the GNW. Thus, on the one hand, we can be presented with an extremely brief stimulus, and still consciously experience it for an extended duration, if it gets amplified and stabilized by top-down signals – once the relevant neural assembly is ignited within the global workspace, the information can remain stable over time, maintaining it "in mind" for as long as needed. On the other hand, even a powerful external stimulus can remain nonconscious if the GNW is distracted elsewhere, thus leading either to complete invisibility (inattentional blindness; Mack and Rock, 1998) or to a transient delay in conscious perception (psychological refractory period; Marti et al., 2012). In a related phenomenon, "retro-cueing", a weak and otherwise invisible stimulus can become conscious, hundreds of milliseconds after its onset, if it is consciously cued (Sergent et al., 2013).

Fourth, given the architecture of the global workspace, we predict that one should be able to decode *any* conscious contents from the workspace neurons, based on the pattern of ignited GNW neurons, which involves a subset of active neurons while the rest are inhibited. In a nutshell, GNW activity holds a neural code for anything that we are conscious of. Notably, GNWT is not merely a theory of conscious perception, but a theory of consciousness in the broader sense. Whether you are conscious of seeing a face, of making an error (Charles et al., 2013) or of being sad, that information must be in the global workspace. Though our field has made substantial progress by focusing on bottom-up visual perception and visual illusions, many other paradigms are available to study other non-perceptual forms of conscious processing (e.g., automatized versus effortful cognitive tasks).

A fifth prediction of the theory refers to the notion of a central bottleneck. The GNW imposes a cognitive bottleneck because it serves as a central core which is shared by all conscious processing tasks, and which can only process once such task at a time. While the workspace is occupied by a specific content, other contents are blocked from entering the workspace, and hence, from accessing consciousness. Those contents may stay at the pre-conscious level, temporarily held in various peripheral buffers, but they are not yet conscious. This idea is supported by several experiments on dual tasking, attentional blink, inattentional blindness, and the psychological refractory period (Marti et al., 2010, Marti et al., 2012, Sergent et al., 2005, Sigman and Dehaene, 2008), where a piece of information has to wait a certain amount of time before becoming conscious. Notably however, the fact that we are always conscious of a single mental object does not imply that this object cannot comprise multiple features. Rather, the theory is compatible with a factorized neuronal representation in which distinct axes of neural space encode different dimensions of knowledge. For instance, in a single moment of conscious access, it is possible to become aware of a person speaking, including multimodal integration of their face, voice, and speech content.

Finally, a sometimes-overlooked property of the workspace model is spontaneous ignition. Conscious ignition does not require external stimulation, but can be also triggered from inside, for instance when accessing a memory or when performing a series of mental calculations. Indeed, simulations show that, even in the absence of external stimuli, the global workspace is constantly traversed by a series of ever-changing global patterns (spontaneous ignitions; Dehaene and Changeux, 2005). This mimics the fact that one can close the eyes and rest while still having a rich conscious flow of information (cf. William James' stream of consciousness; James, 1892). In fact, in patients with disorders of consciousness, the dynamics of spontaneous brain activity can be one of the clearest signatures that the patient is still having a conscious experience. We have made quite some progress in identifying such signatures and using them in the clinic or during anesthetic sedation, as means to detect consciousness (Barttfeld et al., 2015, Demertzi et al., 2019).

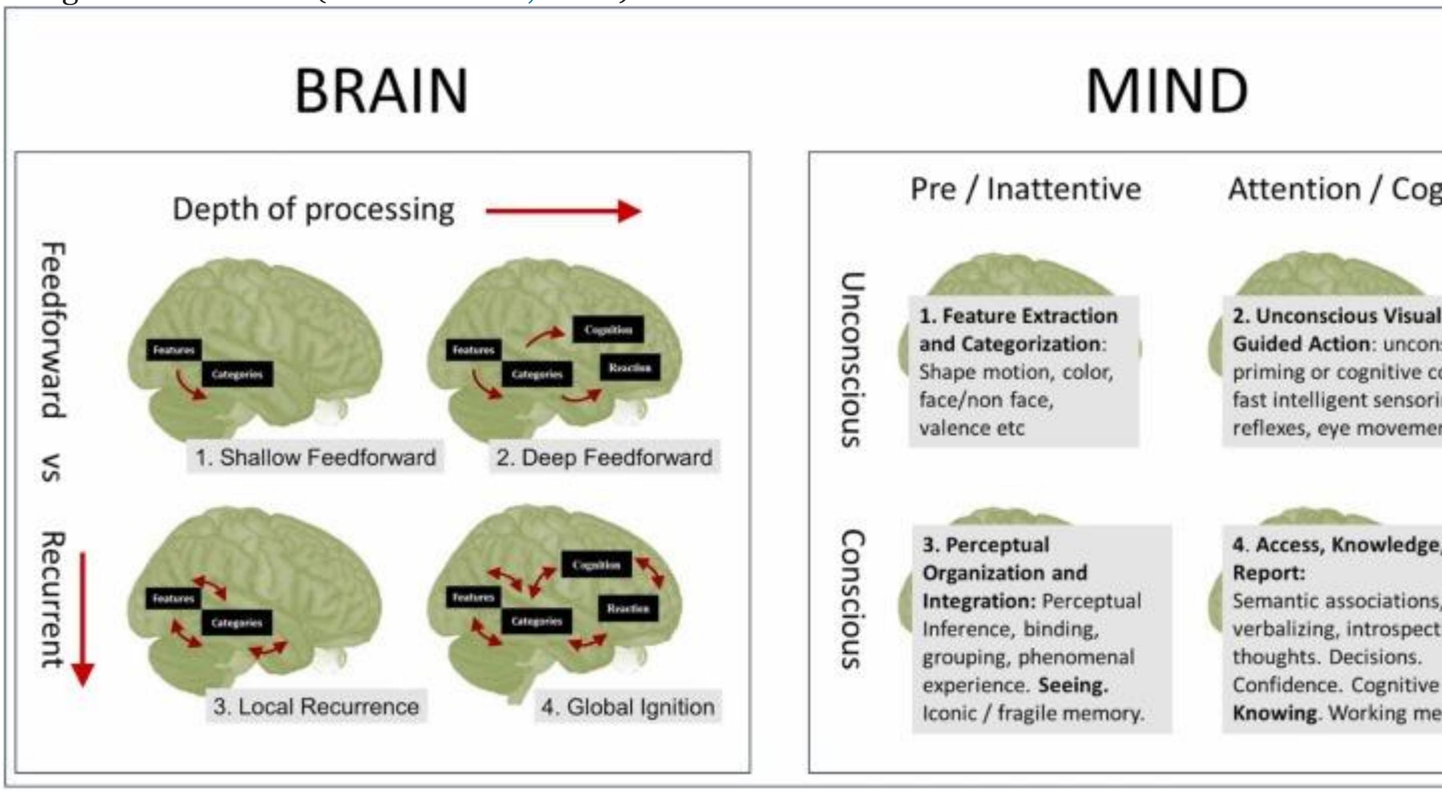
These predictions of GNWT have been supported by experimental observations. For example, Pieter Roelfsema and his team (Van Vugt et al., 2018) simultaneously recorded the activity of V1, V4 and prefrontal neurons in macaque monkeys presented with liminal stimuli – a dot, such that sometimes monkeys saw the dot and sometimes they did not. Areas in the visual cortex were strongly activated both when the monkeys reported seeing the dot and also when they reported not seeing it. This suggests that these areas cannot be the basis of conscious perception as they do not discriminate between conscious and unconscious content. Prefrontal cortex neurons, on the other hand, showed the expected non-linear ignition whenever the monkeys reported seeing the dot. This was true whether it was presented or not (i.e., false alarm). This lends initial evidence for the theory; yet, to provide much stronger support we should go beyond that and show that *any* conscious content has a neural assembly in prefrontal cortex that responds non-linearly to it when it is consciously perceived. This is one of our current research directions. Marie Bellet and other collaborators at Neurospin, including Fanis Panagiotaropoulos, Timo van Kerkoerle, Joaquin and Marie Bellet, and Marion Gay recorded from prefrontal neurons as monkeys were presented with a visual version of the local global test (Bekinschtein et al., 2009, Bellet et al., 2024). Here, we habituated monkeys to a particular type of sequence, such as three identical pictures that are followed by a different one (AAAB). Then, sometimes we presented rare sequences in which all images are identical (AAAA; so they violate the AAAB pattern). Contrary to some popular beliefs, we found that detailed visual information can be decoded from prefrontal cortex (see also Bellet et al., 2022; Panagiotaropoulos et al., 2012). Moreover, we found that whatever the monkey consciously knows, whether abstract or concrete, is encoded by a dimension of neuronal firing in prefrontal cortex. In this experiment, from Utah array recordings in prefrontal cortex (PFC), we could decode which image had been presented (identity), and also what its ordinal position was (number), whether it was different from the previous ones (local effect), which abstract sequence pattern it was included in (global pattern: AAAA or AAAB), and whether this pattern was occasionally violated (global effect). Thus, not only sensory contents, but also abstract knowledge was encoded in PFC firing. In other words, this experiment provides evidence for a key prediction of GNW theory, namely that whatever we consciously know is being encoded by the firing of a distributed population of neurons in prefrontal cortex. Interestingly, these features are encoded by directions in vector space that are almost entirely orthogonal to each other (see Tian et al., 2024; Xie et al., 2022 for relevant work suggesting that there are subspaces in prefrontal cortex for conscious contents). Future work should further test GNWT's prediction that the information encoded by PFC neuronal populations should be just as detailed as the phenomenal content of consciousness – if, for instance, the objective perceptual image is subjectively distorted by a visual illusion, then GNWT predicts that the subjective, rather than the objective information, should be decodable from PFC.

As a final point, I would like to focus on consciousness in humans and other animals. GNW theory proposes that the basic *mechanisms* of conscious access (non-linear ignition and global broadcasting) are the same in human and non-human species (for evidence of a threshold for ignition in crows, see Nieder et al., 2020). The basic operations of conscious access are similar: both have a global workspace and show very similar signatures of consciousness, including spontaneous activity, the presence of visual illusions, central collision etc. However, such parallels in the *mechanisms* of conscious access do not imply that the *contents* of consciousness are identical. There is growing evidence that monkeys may not be able to represent nested compositional structures, unlike humans. In the past five years, I have reoriented some of my research to ask what differentiates monkeys from humans. My hypothesis is that only humans have access to a language of thought: a compositional, syntactic ability which is manifested not only in natural language but also in mathematics or music (Dehaene et al., 2022). While both human and non-human primates share some neural mechanisms for representing sequences, the level of nested symbolic structures may be unique to humans (Dehaene et al., 2015). This would allow us to entertain more complex, symbolic, recursive conscious contents, including the representation of our thoughts and those of others (theory of mind).

2.2. Recurrent processing theory / Victor Lamme

Perhaps the best way to present recurrent processing theory (RPT) is to define it in contrast to other theories and explain where it surpasses these. As a starting point, I would like to note that RPT simply began as a set of empirical observations.
The first empirical observation pertains to the pattern of information processing when a visual stimulus is presented, which involves four stages of processing (Lamme, 2010; see also Fig. 2 below): until about 100 ms post stimulus presentation, feedforward feature extraction of different attributes takes place, like motion, orientation, color and so on (Lamme and Roelfsema, 2000). Notably, at ~200 ms, this feedforward sweep may then proceed up to frontal motor areas, where actions are being prepared, allowing deeper feedforward processing (Lamme, 2018). Third, around the same time, neurons in visual cortex start to engage in local recurrent interactions: higher level neurons that have been activated by feedforward processing send signals back to lower levels, and such repeated back-and-forth interactions enable low and high-level feature extraction to interact. Neurophysiologically, this is accompanied by phenomena like 'contextual modulation', where lower-level neurons modulate their activity depending on the global perceptual context of the features they encode. In this way, recurrent interactions enable functions like perceptual grouping, binding and perceptual organization (Lamme,

2020). Finally, we observe what within GNWT is called 'global ignition', where the whole brain gets engaged into widespread interactions, enabling working memory, recognition and so on (Dehaene et al., 2006).



1. Download: Download high-res image (339KB)
2. Download: Download full-size image

Fig. 2. Recurrent Processing Theory (RPT) identifies four stages of processing on orthogonal axes. The one axis pertains to the depth of cortical processing, or the extent information travels and is processed along the cortical hierarchy of visual areas and towards frontal regions involved in cognition, report and action. This depth is determined by attention; only (bottom-up or top-down) attended stimuli reach the whole depth. The second axis pertains to whether processing is strictly feedforward or involves recurrent interactions (mediated by horizontal connections within areas or feedback from higher to lower areas). The right panel lists the mental / functional equivalents of these four stages of cortical processing. Note how recurrent processing yields two varieties of conscious experience, the one summarized as 'seeing' (Phenomenal -conscious), the other as 'knowing' (Access-conscious), depending on the extent / depth of recurrent interactions. Also note how consciousness is orthogonal to

and independent of attention or cognition, so that in RPT, consciousness gets a proper and fully independent ontology.
The second empirical observation was that the feedforward sweep, no matter how deep it penetrates, is unconscious (Lamme, 2018): It is well-known that one can record both low- and high-level feature responses from anesthetized animals. Studies using manipulations rendering stimuli invisible (like masking) show that feedforward processing may unconsciously activate high-level visual areas (like the Fusiform Face Area (FFA); Fahrenfort et al., 2012), or even frontal regions, thereby evoking unconsciously triggered inhibitory control (van Gaal et al., 2008).
The third empirical observation, on the other hand, was that having conscious sensations invariably goes along with observing recurrent processing: it is present in awake animals, but not anesthetized ones (Lamme et al., 1998), it marks the transition between visible versus invisible stimuli in masking paradigms (Fahrenfort et al., 2007, Lamme et al., 2001), as well as the difference between spontaneously seen or not seen stimuli (Super et al., 2001).
The fourth empirical observation was that as soon as recurrent interactions within visual cortex grow more widespread and include frontal regions, the transition towards attending the stimulus begins, storing it in working memory, or being able to report it (Scholte et al., 2006). One can say: "yeah, I saw that stimulus".
These observations – basically finding that recurrent interactions go along with conscious states and events, while feedforward processing does not - were the empirical foundations for RPT. The main theoretical point that RPT tries to make is very simple: the local recurrence might already be sufficient for us to have a conscious visual percept, to have a phenomenal experience (Lamme, 2018). So that implies that consciousness arises at stage 3 (Fig. 2), rather than at stage 4, where for example GNWT would position it. RPT states that global recurrence might actually only be needed for cognitive access: to be able to report of that conscious experience. In a similar way, there is no need for any higher-order re-representation of the content, as Higher-Order theories hold: first-order representation is sufficient for consciousness to take place (Fig. 2).
This view is very much in line with the empirical agenda of finding the 'true NCC', instead of its consequences, which include reports (Aru et al., 2012). Recently, there have been quite a few experiments suggesting that much of the observed frontal involvement, including the P3b component, are more tightly related to the report and

task performance rather than to the phenomenal experience itself (Pitts et al., 2014, Tsuchiya et al., 2015). Notably, already in the early experiments we did with monkeys, I noticed that recurrent interactions in visual cortex, while related to stimulus saliency and visibility and conscious experience, can easily be dissociated from the report the monkey is making (usually via eye movements) by using different decision criteria (Super et al., 2001). In fact, this exact finding inspired me to posit that recurrent activation in itself is sufficient for having the phenomenal experience, which is different from reporting having that experience. It would imply that local recurrence is independent of attention, access and report (Lamme, 2020), and so may be a better candidate for being a 'true NCC' (Aru et al., 2012).

Yet one could present counter-arguments, mainly because people have different criteria for defining consciousness. Accordingly, GNWT proponents might say that what I refer to as phenomenal experience is in fact, pre-conscious because it precedes access, and because there is no cognition involved. HOT supporters might claim that these states are unconscious, because there is no higher-order thought that points at them. These deep disagreements about the definitions themselves might be one of the reasons for the current stalemate between theories, with 'back of the head' theories – RPT and IIT – claiming that visual activation itself is sufficient for consciousness, and the 'frontal theories' – GNWT and HOT (although there's more to global workspace than just frontal activations) – stating that frontal activations are necessary for consciousness to take place. In a way, this entire debate becomes almost a matter of taste, hinging on which criterion you use to define consciousness.

So, how can we find a way out of here? I believe a promising attitude is to focus on what needs to be explained about conscious vision. First, a very obvious and prominent hallmark of conscious visual perception, and all sorts of perception, is that experience is unified. Although there are about 30–40 different visual areas in the brain, each processing a different type of visual information, we have one visual percept: we don't see its aspects (e.g., shapes, colors or objects) in isolation, but as an integrated whole. Second, conscious vision is a type of inference. We usually go beyond the physical information of the stimulus itself to what we interpret based on that information. We go from luminance to gray shade, from wavelength to color, from features to inference. Third, consciousness is generally cognitively impenetrable. Even if we gain cognitive information about a specific illusion, we cannot but see the illusory percept. Fourth, consciousness involves integration. When we see a face, there are neural mechanisms (e.g., face neurons, or the FFA) that detect the difference between faces and other stimuli, seeing a face is way more than that: we integrate this 'faceness property' with the shape of the face, its color, its structure and all sorts of other properties it has. Thus, seeing involves a large degree of integration. To me, this is a key feature that we must explain about consciousness. In a way, this is the explanandum.

This point is nicely illustrated by a study conducted by Johannes Fahrenfort (Fahrenfort et al., 2012), showing that the FFA is activated by invisible faces, and that only when the FFA interacts with lower visual areas, you get a conscious percept of that face. I claim that local recurrence itself is fully capable of explaining this unity of consciousness, or this integration, because such local recurrence mediates those functions: perceptual inference, incremental Gestalt grouping, long-range and complex organizations, figure-ground organization etc. (Lamme, 2014). I really don't see how any frontal or higher order theories have any explanatory power towards explaining this aspect of consciousness - the aspect of unity, integration and binding. The same goes for cognitive impenetrability: I argue that frontal theories would predict that it should be entirely possible for knowledge to affect perception (e.g., to overcome perceptual illusions). If all types of processing get integrated into one coherent whole by this global ignition, why wouldn't you be able to also integrate your thoughts into the perception? I believe that frontal theories would also predict dependence of perceptual functions on attention, but that is not the case: there is ample evidence that all these above-mentioned functions happen whether you attend to a stimulus or not. Attention may modulate or strengthen these functions, but they do not critically depend on it (Lamme, 2020). This has been demonstrated by yet another nice experiment of Johannes Fahrenfort (Fahrenfort et al., 2017), where the effects of attentional blink (AB) and masking on processing Kanizsa figures were compared. He decoded the contrast (high vs. low) of Kanizsa inducers from visual cortex activity regardless of AB or masking, showing contrast is a feature typically processed by feedforward processing (Fahrenfort et al., 2017). Conversely, the 'integration contrast' (i.e., whether the Kanizsa configuration yielded the typical illusory percept) was disrupted by masking (known to interfere with visibility and recurrent interactions in visual cortex), yet not by the attentional blink.

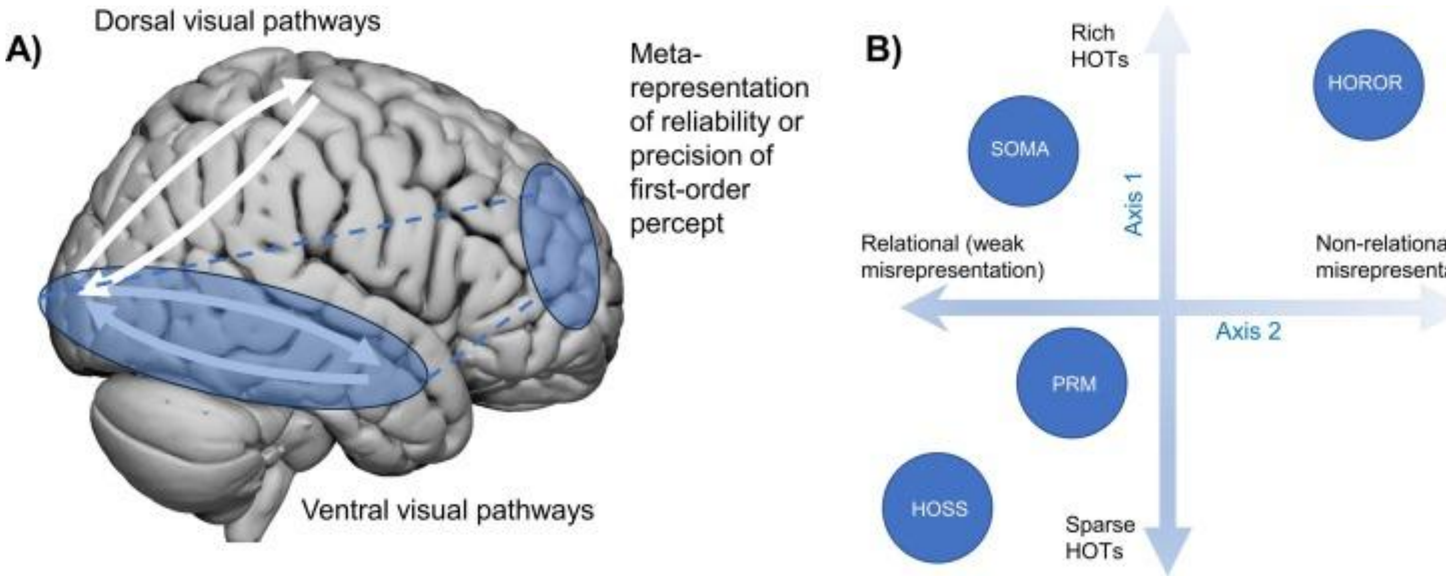
Another feature of visual perception is that it is phenomenally very rich. This poses a problem for frontal theories, which are so tightly linked to attention, having strong capacity limits. Thus, these theories should in fact predict that visual perception is fairly limited and sparse, having an attentional bottleneck capacity. Now, one can claim, for example, that change blindness experiments show exactly that: if you show an array of objects twice, and in the second presentation you change one of the items, it is indeed very difficult to notice the difference between the two. Yet we have shown in many experiments that if you cue these arrays between the change, participants clearly recognize the item that was replaced (Landman et al., 2003).

This suggests that their representation of the first image is in fact very detailed, precise and rich (Sligte et al., 2010). These iconic and fragile memory experiments have clearly shown that in change blindness, the first scene is overwritten by the second scene as soon as it is presented. Accordingly, change blindness is a failure of memory, rather than of perception (Lamme, 2010).

As a bonus argument for RPT, I will take a metaphysical standpoint: I think that RPT is ideally suited to give consciousness an independent ontology. The four stages of processing described above can be put in a 2 × 2 scheme where attention or cognition can be seen as one axis, basically depending on how deep information gets processed in the brain, whether frontal regions are involved etc. The other axis differentiates between unconscious and conscious processing, which is an entirely different and orthogonal axis, and in the brain corresponds to the difference between feedforward and recurrent processing. This scheme readily explains how shallow versus deep feedforward processing makes it possible to have unconscious priming depend on attention (Naccache et al., 2002), or why strictly feedforward frontal activation may yield unconscious cognitive control (van Gaal and Lamme, 2012). Similarly, it explains the difference between phenomenal and access consciousness (Block, 1995), indexing 'shallow' (visual) versus 'deep' (global ignition) recurrent processing. Most importantly, it shows how in RPT consciousness is thought to be independent and orthogonal to other cognitive functions such as attention or cognition, functions that are often conflated with consciousness in frontal theories like GNWT or HOT. One could say that RPT is a theory that takes consciousness seriously, and gives it its own and independent ontological status.

2.3. Higher Order theories / Stephen Fleming

Higher Order theories (HOTs) are already well-established in philosophy (Carruthers, 2001), but are somewhat more of a new kid on the block when it comes to neuroscience. As previous work shows (Yaron et al., 2022), there is less empirical evidence for or against these theories compared to the others. But I am still excited by HOTs, because I think they offer us a way of charting an empirically productive path between global and local theories. I will try to demonstrate that using the guiding questions of this debate.
First, what is the explanandum? We are trying to explain the presence versus absence of phenomenal experience: for example, if our visual systems are processing information about a sunset over London, we are usually aware of, and able to communicate, our experience of the sunset to others, while remaining unaware of other perceptual inputs that are nevertheless being processed, such as the feeling of clothes on our skin, changes in our posture and so on. HOTs seek to explain the difference between these two types of information processing.
Second, what do HOTs claim? In a nutshell, HOTs say that a perceptual representation of some content X – for instance, a red apple – is not sufficient for a conscious experience of X to arise. These kinds of representations are known as "first-order" as they are directed at the world. Having a first-order representation of the object is often crucial for guiding behavior – for instance, allowing us to pick up and eat the apple. However, according to HOTs, such first-order representations can also occur nonconsciously and are insufficient for phenomenally conscious experiences. Instead, HOTs hold that the phenomenal consciousness of X depends on an organism being in some way aware of being in state X. This, in turn, entails that the first-order state is in some way monitored or meta-represented by a higher order representation (Brown et al., 2019, Lau and Rosenthal, 2011; Fig. 3). This higher-order representation can take many forms, as described below. HOTs account for unconscious influences on behavior by positing that first-order states can still drive task performance unconsciously. In particular, those first-order states might be quite widely broadcast and facilitate unconscious control – which is where HOTs start to diverge from GNW.



1. Download: Download high-res image (258KB)
2. Download: Download full-size image

Fig. 3. **a.** Illustration of hypothetical neural substrates supporting lean higher-order accounts such as HOSS. Prefrontal areas support implicit meta-representations of the reliability, or confidence in, first-order percepts built along ventral visual pathways. The meta-representation is hypothesized to track the reliability of an entire first-order representation, which itself may be supported by hierarchical processing along ventral visual pathways. Under HO theories, global broadcast and action guidance may still proceed unconsciously, for instance via the dorsal visual pathways. **b.** Two axes of disagreement between higher-order theory variants: HOSS, Higher-Order State Space account; PRM, Perceptual Reality Monitoring account; SOMA, Self-Organizing Metarepresentational Account; HOROR, Higher-Order Representation of a Representation account. A first axis indicates to what degree meta-representations are rich or sparse. Sparse HO theories propose that phenomenal experience is jointly determined by first-order (FO) and HO states, with HO states playing a lean role in tracking the precision, intensity or reliability of FO states. Rich HO theories propose that phenomenal experience is fully determined by higher-order states, such that higher-order representations are just as rich and detailed as perceptual experience itself. A second axis of disagreement is on whether HO representations can “misrepresent” their FO targets, and in what way. The position of theory variants along these axes is often correlated, with the rich non-relational theories permitting stronger misrepresentation.

Importantly, there are multiple versions of HOT, just like there are multiple versions of first-order theories. We are actively engaged in an adversarial collaboration to test distinct empirical predictions made by the different HOTs, led by myself and Axel Cleeremans and supported by the Templeton World Charity Foundation. This project is seeking to test two key axes of disagreement between different HOTs (Fig. 3b). One key difference is whether a higher-order theory allows the first-order states to be doing any of the work involved in creating conscious experience. On the one hand, proponents of "sparse" views suggest that consciousness arises when first-order and higher-order states work together (also referred to as joint determination; Fleming, 2020; Lau, 2019). According to these views, the content of experience is carried by the first-order states, with the higher-order states having a more subtle (but nevertheless critical) role in monitoring the precision, intensity or reliability of the first-order state. These types of higher-order representation are sometimes referred to as "pointers" or "indexes". At the other end of the spectrum, we have "rich" HOTs that predict that the phenomenal experience is fully determined by the higher-order states, which are held to be as rich as perception itself (Brown, 2015, Rosenthal, 2005). Irrespective of the question of perception being sparse or rich, these variants hold that the higher-order states should fully determine the sparse or rich content of conscious perception. Another axis of disagreement concerns whether different variants of HOT allow for the possibility of misrepresentation of first-order states – cases in which the higher-order and first-order representations disagree. Under relational views (which are also often rich; Fig. 3b) what matters for consciousness is the content of the higher-order, not first-order states – and so consciousness should change in tandem with the former and not the latter. There are also intermediate views, in which the first-order states contribute to conscious experience, but the higher-order states are more than mere pointers or indexes, and also furnish independent content (Cleeremans et al., 2020, LeDoux, 2020a, Van Gulick, 2004).

HOTs also hold that it is particularly important to control experimentally for performance when adjudicating between different theories of consciousness. Plants, cameras and thermometers all have some sensitivity (or, in signal detection terminology, some level of d') to aspects of the environment, yet we typically do not think they are conscious of what they are sensitive to. Similarly, in the human brain, there may be mechanisms that create a form of behavioral sensitivity to the stimulus, without any accompanying conscious experience. This creates an empirical problem in human experiments, because if first-order representations can both drive performance and contribute to consciousness, then if we alter consciousness by also changing performance or signal strength, we will not know whether to ascribe the effects we find to performance or to consciousness. This is a pervasive problem for the field, but it is particularly problematic for testing predictions of HOTs, because if our experiments do not control for performance, we might unfairly stack the deck in favor of first-order theories by misattributing a change in consciousness to the accompanying change in first-order states that may in fact be (unconsciously) driving performance (see discussion in Lau, 2022).

HOTs are also often linked to metacognition – the ability to reflect on and monitor other aspects of cognition and perception (Fleming, 2024). Metacognition is often investigated by asking people to judge confidence (or error) in aspects of behavioral and cognitive performance – which in perceptual tasks is referred to as "perceptual confidence". Because in studies of metacognition, people are usually explicitly asked to reflect on and judge their performance, these experiments are tapping into "explicit" forms of metacognition. This can often be a source of misunderstanding. Higher-order theorists do not typically hold that explicit metacognition is the same thing as phenomenal consciousness. Instead, some variants of HOT claim that there are theoretical reasons to think that the computations that are important for instantiating higher-order states, such as monitoring the precision of first-order representations, share mechanisms with perceptual confidence (Fleming, 2020, Lau, 2019). This is because both may depend on a subpersonal implicit form of metacognition – aspects of metacognition which are not necessarily available for subjective report.

Empirically, studying perceptual metacognition is useful for two reasons: first, it allows us to study the neural basis of perceptual confidence, and second, it allows us to tackle the performance issue, by using statistics such as metacognitive efficiency which isolate metacognitive capacities controlling for first-order task performance (Maniscalco and Lau, 2012). By analyzing metacognitive efficiency in experiments probing conscious experience, we can identify situations where performance (d') is matched, but there is an isolated change in confidence. In the past few years, a growing body of work has implicated prefrontal cortex, and particularly anterior prefrontal sub-regions, in perceptual metacognition. In our lab, we find early implicit signatures of perceptual confidence tracked in the activity of the medial prefrontal cortex, particularly the perigenual anterior cingulate cortex, followed by a later involvement of the lateral frontopolar cortex (Bang et al., 2020, Bang and Fleming, 2018). This division of labor between medial and lateral subregions of prefrontal cortex in perceptual metacognition has also been shown by Marios Philiastides' group using EEG-informed fMRI (Gherman and Philiastides, 2018), and in studies of the role of intra-prefrontal connectivity in supporting metacognitive efficiency (De Martino et al., 2013). One interpretation of these data is that the coupling supports a more explicit form of metacognition, so that confidence estimates become available for communication and control (Fleming, 2024). Other studies have found that when activity within subregions of the prefrontal cortex is manipulated using Transcranial Magnetic Stimulation (Shekhar and Rahnev, 2018) or decoded neurofeedback (Cortese et al., 2020) it is possible to modulate perceptual confidence without changing performance. These data support the notion that prefrontal and parietal regions are important for perceptual metacognition.

A key implication of these findings is that there is a psychologically and neurally meaningful distinction between implicit and explicit metacognition: perceptual confidence can be formed automatically, independent of the ability to communicate metacognitive estimates to others (Fleming, 2024, Shea et al., 2014). Studying the role of implicit confidence estimation in perceptual experience will be helpful for adjudicating between the different variants of HOT. But even though HOTs traditionally emphasize prefrontal cortex as supporting higher-order states, there is increasing recognition that the critical distinction here is computational rather than anatomical. As Joseph LeDoux has pointed out, mapping a philosophical division between first-order and higher-order states onto the functional anatomy of the human brain is necessarily going to be much more complicated than two interacting neural populations (LeDoux, 2020b). Thus, the 'front versus back of the brain debate' (e.g., Boly et al., 2017; Odegaard et al., 2017) should be viewed as more of a starting point than endpoint for adjudicating between the different theories.

In that vein, I want to briefly highlight a research program that we have been pursuing, developing a minimal computational implementation of a sparse HOT within a predictive coding framework. We call this the higher-order state space (HOSS) model (Fleming, 2020). We start with a first-order generative model that performs inference on perceptual inputs, such as whether a visual stimulus is an apple or an orange. This process of building first-order representations of the world may proceed unconsciously, as in Helmholtz's famous phrase of "unconscious inference" (Gregory, 1970, Helmholtz, 1867/1962) – we perceive the product of perceptual inference, rather than the machinery of inference itself. Within HOSS, additional higher-order layers monitor the precision of these first-order representations to allow the system to track whether the first-order representations reflect external reality, or just noise (similar architectures underpin the perceptual reality monitoring variant of HOT; Lau, 2019). The operation of these higher-order representations underpins conscious experience by endowing the agent with beliefs about the assertoric force and reality of its perceptual representations – beliefs which have what philosophers refer to as "assertoric force" (Lau, 2022).

These different types of computation make broad qualitative predictions about the nature of the neural representations at these different levels: higher-order states should be lower dimensional and symmetrically encode the high vs. low precision of the first-order states, whereas first-order states should be higher dimensional and encode first-order content (note that a first-order state may be supported by an anatomical network that spans sensory and association cortices, rather than a single node in early sensory areas). We have also been exploring how classical neural signatures associated with conscious awareness such as global ignition can be reinterpreted within this model architecture: specifically, we see that ignition-like signatures emerge as a natural consequence of an asymmetry in the first-order state prediction errors. When the representations within a first-order generative model are precise (on trials associated with conscious awareness of stimulus contents), strong prediction errors from the stimulus propagate through the system (Whyte et al., 2022). Thus, predictive processing versions of HOT may be able to recapitulate the neural signatures previously highlighted as supporting GNWT. However, HOSS interprets these ignition-like signatures differently: rather than reflecting global broadcast, under HOSS, conscious trials are associated with greater (average) prediction error. This is due to an asymmetry in the magnitude of prediction errors associated with "aware" and "unaware" trials in an experiment. This interplay between predictions and prediction errors is consistent with the broader class of predictive processing (PP) models discussed earlier, but now focused on explaining differences that ensue between conscious and unconscious perception.

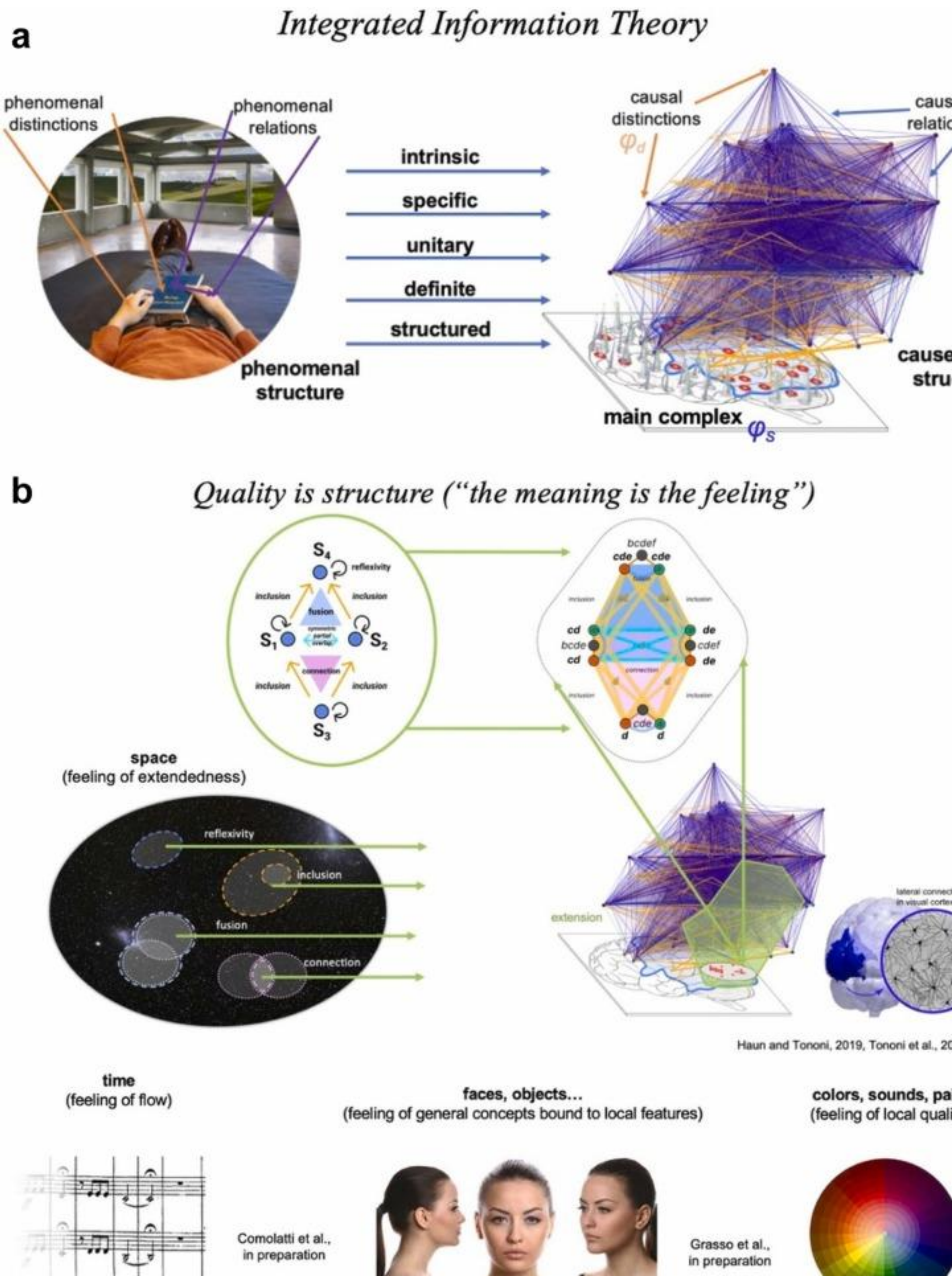
In work together with Nadine Dijkstra, Peter Kok and colleagues at UCL we have started to test some of these predictions of the model using neuroimaging. We trained participants to generate predictions not only about the content of the stimulus (first-order) but also whether or not they will be aware of that content (higher-order). For instance, on some trials, you might be cued to expect that you will not see anything (an expectation of stimulus absence), but that if you did see something, the stimulus would be a face (as opposed to a house). By confirming and violating these content and awareness expectations with face and house stimuli embedded in noise, simulations show that we can nicely orthogonalize the prediction errors at the two levels (first-order and higher-order) of our computational model. We find that behaviorally, these two types of prediction error both affect response times, slowing down identification of the stimulus (Dijkstra et al., 2024), and modulate subjective reports and confidence in stimulus identity (Haarsma et al., 2024). We then combined a no-report version of this experimental design with fMRI: participants were simply presented with the cues and stimuli without requiring overt responses. In the brain, we find that while first-order state prediction errors are tracked in patterns of activity in ventral visual cortical areas, higher-order state prediction errors are tracked by activity patterns in medial PFC (Dijkstra et al., 2023). Taken together, these findings support the dissociation in prediction errors posited by the HOSS model.
Finally, I would like to end by highlighting common ground between some of the theories under consideration, which have more in common than is typically assumed: HOTs and GNWT (see section by Dehaene above) both claim that consciousness depends on interactions between first-order states and higher-order computational processes, and under the predictive coding implementations of HOT that we have been pursuing, ignition and broadcast signatures reflect asymmetries in global prediction error as a function of precision. HOTs and RPT (see section by Lamme above) agree on some things too: recurrent message passing may be needed for higher-order and first-order states to interact, so this approach is especially compatible with the joint determination variants of HOT. There is obvious common ground with PP theories (see section by Seth below): as we have seen, it is possible to implement versions of HOT within generative model architectures. Under PP there seems to be no single criterion by which conscious and unconscious contents are differentiated; HOSS, however, provides a clear proposal (see above). Lastly, with respect to IIT (see section by Boly below): in a

recent review (Lau et al., 2022) we suggested that the organization and structure of the first-order states themselves – the sensory code, the fact that it's smooth and sparse – may enable higher-order representations to support the relational comparisons that might underlie the ability to communicate what it is like to have an experience: the capacity to implicitly represent that scarlet is more similar to crimson than it is to blue, for instance. To me, this is what a functional explanation of the "what it is likeness" of conscious experience may look like, and it has some commonalities with the idea that the organization and cause-effect structure of the first-order states themselves are important for phenomenology, as proposed by IIT (although under HOT, this organization alone would not be sufficient for conscious experience). Thus, different elements within HOTs share common ground with other theories of consciousness.

2.4. Integrated Information Theory / Melanie Boly

A theory of consciousness must answer two key questions. First, what determines whether consciousness is present versus absent; and second, what determines why specific experiences feel the way they do. I will start with the first question. With respect to anatomy, why do certain parts of the corticothalamic system seem to contribute directly to consciousness, while many other parts of the brain, such as the cerebellum and certain cortical areas, do not (Tononi et al., 2016)? With respect to physiology, why is it that consciousness vanishes during deep sleep even though neurons continue to be active, and during generalized tonic-clonic seizures, even though neurons fire maximally and in a highly synchronous manner? (Juan et al., 2023).
The second question a theory of consciousness should answer is what determines why specific experiences feel the way they do. This includes explaining why visual space or body space feel extended, why time feels flowing, and why objects, colors, sounds or touch, feel the way they do.
IIT does not ask how the physical world 'gives rise' to experience—it does not try to 'squeeze' consciousness out of the gray matter of the brain. Instead, IIT starts from consciousness itself—from phenomenology—which is the starting point for everything, including science. IIT's approach is to characterize the essential properties of consciousness—those that are immediately and irrefutably true of every conceivable experience—and formulate a principled, coherent account of those properties based on

the causal powers of a physical substrate (Fig. 4; see also Albantakis et al., 2023; Tononi et al., 2016).


a
Integrated Information Theory
phenomenal distinctions
phenomenal relations
intrinsic
specific
unitary
definite
structured
phenomenal structure
causal distinctions
φ_d
main complex
φ_s
b
Quality is structure ("the meaning is the feeling")
S_4
reflexivity
inclusion
fusion
S_1
S_2
connection
S_3
bcdef
cde
cd
bcde
de
cdef
d
space
(feeling of extendedness)
reflexivity
inclusion
fusion
connection
extension
Haun and Tononi, 2019, Tononi et al.,
time
(feeling of flow)
faces, objects...
(feeling of general concepts bound to local features)
Comolatti et al., in preparation
Grasso et al., in preparation

1. Download: Download high-res image (783KB)
2. Download: Download full-size image

Fig. 4. **a.** The starting point of IIT are the essential properties of consciousness – those that are true of every experience. In short, every experience is intrinsic, specific, unitary, definite, and structured by phenomenal distinctions bound by relations (left side). IIT then formulates these properties in physical terms, understood operationally as cause-effect power, or the ability to take and make a difference. Accordingly, the substrate of consciousness, or main complex, must have cause-effect power that is intrinsic, specific, irreducible, definite, and structured (right side). The cause-effect structure fully unfolds the complex's cause-effect by considering causal distinctions, i.e., the cause-effect of each of its subsets, and causal relations, i.e., the overlaps among cause-effects. The content of experience should be accounted for in full by the complex's cause-effect structure, with no additional ingredients. **b.** Accordingly, for IIT all quality is structure. Moreover, the meaning of contents of experience is fully intrinsic, being specified by the cause-effect structure ("the meaning is the feeling"). IIT's research program aims to account for the quality of different kinds of experiences. As a first example, an analysis of spatial experiences indicates that space feels "extended" because phenomenal distinctions (called "spots") are bound by relations that capture reflexivity, inclusion, connection, and fusion. An analysis of cause-effect structures unfolded from substrates connected in a grid-like manner, as is the case in much of posterior cortex, can account for phenomenal extendedness in physical terms. Ongoing work aims at accounting for why time feels flowing, why faces and objects feel like general concepts bound to local features, and why colors and sounds feel the way they do. IIT's explanatory identity claims that a complex's cause-effect structure should account for all properties of an experience, essential and accidental, with no additional ingredients. Note that the explanatory identity is between the quality (structure) of an experience and the cause-effect structure (composed of causal distinctions and relations) specified by the complex in its current state; not with the complex's connectivity or activity patterns. The cause-effect structure of the complex is obtained by unfolding the cause-effect power of all its subsets in their current state, following IIT's composition postulate.

What do we mean by physical? IIT defines the physical in purely operational terms, as cause-effect power—the ability of a substrate's units to take and make a difference. In line with the traditional scientific method, it employs systematic manipulations and observations of a physical substrate to obtain a Transition Probability Matrix (TPM) that summarizes the substrate's powers in terms of causes and effects: if we do X, what is the probability that we reliably observe Y (Albantakis et al., 2023).

IIT identifies five properties that hold true for every conceivable experience (Albantakis et al., 2023). These are intrinsicality (experience exists for itself), information (it is specific), integration (it is irreducible), exclusion (it is definite), and composition (it is structured). The last property (composition) refers to the fact that every experience is composed of phenomenal distinctions – such as faces, the left and right corner of space and so on – and phenomenal relations that bind them together in various ways – for example, the face is in the left corner.

This phenomenal structure composed of distinctions and relations now needs to be accounted for in physical terms. IIT does so by developing a mathematical framework that allows us to 'unfold' in full the cause-effect power of a substrate, leading to a cause-effect structure (also called Φ structure) composed of causal distinctions (cause-effects) and relations (overlaps among causes and/or effects). According to IIT, this is the physical structure that can account for all the properties of consciousness, and for the quality of a specific experience, here and now.

Within this mathematical framework, which we have refined over the years, we can aim to achieve an explanatory identity such that all of the properties of experience can be expressed explicitly in terms of a specific Φ-structure expressing a substrate's cause-effect power (Albantakis et al., 2023). In this way, we can generate predictions about which physical substrates can support consciousness (and which cannot), as well as about the quality of experience (the composition of the Φ-structure) and its quantity (measured by Φ). Importantly, given a substrate in its current state, the associated Φ-structure should account for the properties of specific experiences with no additional ingredients.

What does the theory explain empirically? With respect to anatomy, IIT explains why certain substrates, and not others, can account for the essential properties of consciousness. For example, most of posterior-central cerebral cortex is organized roughly like a hierarchy of 2D grids of neurons (i.e., pyramid of grids; Maruoka et al., 2017). According to IIT, this kind of substrate is very well suited for supporting rich cause-effect structures of high Φ. In contrast, the modular architecture of the cerebellum (D'Angelo and Casali, 2013) necessarily fragments into many small substrates, each supporting disjoint cause-effect structures of minimal Φ. With respect to physiology, IIT can explain why the same anatomical substrate – the cerebral cortex – can support large Φ-structures during wakefulness, but disintegrates into many small structures during dreamless non-REM sleep, when changes in neuromodulation lead to the breakdown of cortical effective connectivity. Indeed, after direct electrical stimulation of the cortex, intracranial recordings show complex, recurrent interactions during wakefulness, which are interrupted by the stereotypical occurrence of an off-state during deep non-REM sleep (Pigorini et al., 2015).

IIT's principles have also inspired a Transcranial Magnetic Stimulation - Electroencephalography (TMS-EEG) method for assessing the presence or absence of consciousness through a Perturbational Complexity Index (PCI): in wakefulness, the EEG response to TMS is complex (high PCI), revealing the induction of differentiated activity patterns, while in deep non-REM sleep the response becomes local, short, and stereotyped (Massimini et al., 2005, Tononi et al., in press). Notably, PCI is high in states where subjects are conscious but unresponsive, including REM sleep or ketamine anesthesia (Casarotto et al., 2016).

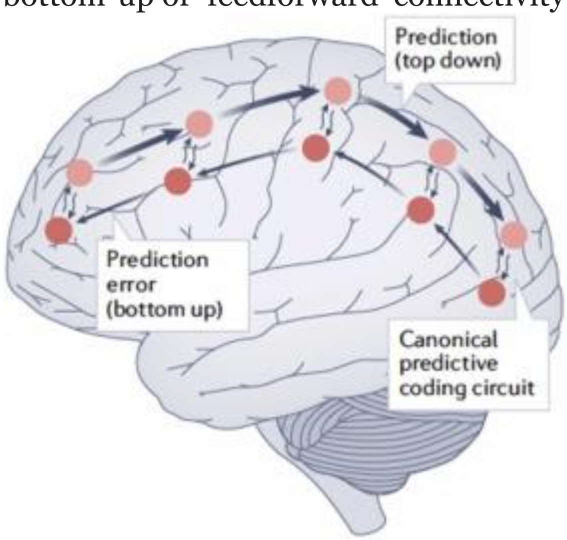
Going back to the question of what determines why specific experiences feel the way they do, we started approaching this question with spatial extendedness, because the experience of space is both pervasive and partially penetrable, in the sense that we can use introspection—especially spatial attention—to dissect its phenomenal structure, as opposed to color or pain (Haun and Tononi, 2019). We then showed that the kind of Φ-structures specified by 2D grids can account for how space feels. Briefly, the fundamental property of spatial experience is extendedness: the visual field, for example, is composed of phenomenal distinctions ('spots') that overlap according to a distinctive set of phenomenal relations (reflexivity, inclusion, connection, and fusion). It turns out that the Φ-structures specified by 2D grids, such as those found in much of posterior-central cortex (Wang et al., 2015), are composed of causal distinctions and relations of exactly the same kind. Given that experience is pervasively spatial (in both the visual and somatosensory domains), it is no surprise that both in humans and in non-human animals, large parts of the posterior-central cortex indeed constitute of pyramids of grid structures. Ongoing work (including an adversarial collaboration, https://osf.io/4rn85) is testing some of the predictions that follow from IIT's account of spatial experience and its neural substrate. In the meantime, we are actively pursuing a research program that aims at accounting for the feeling of time flowing (Comolatti et al., 2024), and of objects binding general concepts with particular features (for more details, see IIT WIKI).

While IIT focuses on the contents of an experience triggered by a stimulus or occurring during a dream, it is obviously meant to deal with both how we perceive the external world and how we learn about it (for an in-depth discussion of how IIT formalism can reconceptualize perception as the triggering of intrinsic meanings/feelings by stimuli, see Mayner et al., 2024). Through the notion of matching, IIT can also explain the extent to which intrinsic feelings/meanings capture regularities due to causal processes in the environment, which are internalized in the connectivity of the brain. Finally, there are some simple demonstrations using animats (small simulated organisms evolving in a simple environment) showing that learning about a complex environment is typically associated with an increase in integrated information (Albantakis et al., 2014), thus providing an indication as to why selective pressure might favor the evolution of greater and greater consciousness. Accordingly, IIT reconceptualizes perception as interpretation, rather than as information processing or representation.

2.5. Predictive processing / Anil Seth

Predictive processing (PP) theories occupy a unique region in the space of theories of consciousness. This is mainly because they are, in general, not (yet) 'theories of consciousness' in the sense of proposing necessary and sufficient conditions for conscious experience (or specific conscious contents) to occur. Instead, they are more general theories of brain and mind that can be applied to account for various properties of consciousness – such as the specific phenomenological properties associated with different kinds of conscious experience. As Howhy & Seth (2020) put it, PP is not a theory *of* consciousness, it is – or least started out as – a theory *for* consciousness science.

According to PP, the brain is continually minimizing sensory "prediction error" signals, either by updating its predictions about the causes of sensory signals or by performing actions to bring about predicted or desired sensory inputs (the latter process being called "active inference"; Clark, 2013; Friston et al., 2010). This ongoing process of prediction error minimization provides a mechanism by which the view of perception as a process of Bayesian inference (Helmholtz, 1867/1962) can be implemented. In this view, the aim of perception is to infer the most likely causes of sensory signals (the Bayesian posterior, or 'best guess'), given some 'prior' belief or expectation about these causes, and the new information provided by the sensory data (the Bayesian likelihood). Importantly, this process can also provide a means of predictive regulation of physiological variables, in which interoceptive perceptual priors can serve as set-points or target ranges for homeostasis and allostasis (Barrett and Simmons, 2015, Seth, 2015). In its most ambitious and all-encompassing version, the "free energy principle" (FEP), the mechanism of prediction error minimization arises out of fundamental constraints regarding control and regulation. These constraints apply to all physical systems that maintain their organization over time in the face of external perturbations, with living systems being a particularly expressive example (Friston et al., 2010). In this view, prediction error emerges as a proxy for sensory entropy, which organisms are mandated to minimize, through active inference, in order to remain in the statistically expected states that are compatible with their survival. While the FEP, being a principle, is not itself testable, 'process theories' that fall within the FEP – such as PP – are testable. Generally, in PP, predictions are proposed to flow in a top-down (or 'inside-out') manner, with prediction error signals flowing in a bottom-up (or 'outside-in') direction (Fig. 5). This challenges the usual ascription of the labels 'feedback' and 'feedforward' to top-down and bottom-up connections respectively, since 'feedback' usually connotes transmission of an error signal, whereas in PP these signals are typically associated with bottom-up or 'feedforward' connectivity.



1. Download: Download high-res image (91KB)

2. Download: Download full-size image

Fig. 5. Predictive processing (and related theories like active inference; Friston et al., 2010) are not *prima facie* theories of consciousness (Hohwy and Seth, 2020). They are more general theories of perception, cognition, and action that can be used to derive predictions relating to aspects of consciousness, perhaps leading to some minimal theory of consciousness *per se* (Whyte et al., 2024). The core claim of predictive processing is that conscious mental states are associated with top-down signaling (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. Adapted from Seth and Bayne (2022).

This functional architecture for PP is not fixed. Tscshantz et al. (2023) recently described a 'hybrid predictive coding' architecture in which predictions and prediction errors flow in both directions, but at different time scales, implementing a flexible balance between learned (amortized) mappings from sensory data to Bayesian posteriors and standard (iterative) inference. Under the FEP, recent extensions have illustrated how minimization of 'expected free energy' (i.e., approximate future prediction error under some action policy) can optimally balance a trade-off between exploratory (epistemic) and goal-directed actions (Friston et al., 2015, Tschantz et al., 2020). I mention these extensions to underline the richness of the computational resource provided by the still-evolving PP (and, more broadly, active inference) framework. This resource is further enriched by the hierarchical nature of PP, and – in particular – by the key role of *precision-weighting*, in which sensory prediction error signals with high (estimated) precision have greater influence on updating predictions; a mechanism that has been argued to map onto the role of attention in perception (Feldman and Friston, 2010).

The broad strategy here is to leverage the rich resources of PP to develop systematic mappings between properties of conscious perception and properties of the underlying neural architectures – where by 'systematic' I mean being systematically guided by theoretical considerations of some sort (Hohwy and Seth, 2020). In virtue of this systematicity, such mappings should also have explanatory and predictive power that can be experimentally tested. The larger ambition is that, through iterating this process, and by examining shared features across diverse individual phenomena, a core set of theoretical commitments will emerge. This set will constitute a PP theory of consciousness *as such*, rather than merely being a theory 'for' consciousness science (Whyte et al., 2024).

Space limitations preclude a detailed review of all the ways this iterative refinement of PP is underway (see Hohwy and Seth, 2020; Seth and Bayne, 2022; Whyte et al., 2022). I will highlight just a few, to illustrate both breadth and depth. An emerging theme in the more recent models is their emphasis on *action* as crucial in shaping conscious contents and transitions between contents, where these actions can be either overt (i.e., expressed via the body) or covert (i.e., mental actions, such as paying attention).

Perhaps the most well-developed PP studies of conscious perception focus on visual experience. An early model tackled binocular rivalry by proposing the existence of two competing perceptual hypotheses, one of which 'wins', leading to perceptual dominance (Hohwy et al., 2008). Sensory signals from the alternative hypothesis accumulate as prediction error, which eventually leads to a perceptual transition, at which the source of unexplained prediction error switches and the cycle repeats. Parr et al. (2019) extended this model, using expected free energy, to propose that the accumulation of uncertainty makes the suppressed sensory signals progressively more epistemically attractive, until attention (considered as a covert action) shifts to these signals, which makes the corresponding perceptual belief sufficiently precise that it dominates perception. Notably, this model accounts for experimentally observed features of binocular rivalry which are hard to account for in alternative 'passive' models: these include the slowing of rivalry in the absence of attention, the modulation of dominance durations by reward, and the regularities relating stimulus features such as luminance contrast with dominance duration (Whyte et al., 2022).

Versions of this model have also been adapted to account for Troxler fading, where peripheral visual content fades from experience (Parr et al., 2019), and to bistable perception of Necker cube stimuli (Novicky et al., 2024), both again emphasizing the idea of policy selection to minimize expected free energy, with perceptual switches being primarily driven by epistemic actions.

The PP approach also has the potential to integrate experimental findings which may challenge some existing theories. In one example using computational modeling, the researchers developed a PP model of conscious access in which the gating of working memory is treated – similarly to attention – as a covert cognitive action (Whyte et al., 2022). Applying this model to a simulated visual masking task, they showed how late P3b-like event-related potentials (ERPs) and increased PFC activity could be induced by the working memory demands of report generation. But when the model is modified to simulate a no-report condition, these (simulated) late ERPs and PFC activities diminish or go away entirely - an empirical observation that has often been taken to challenge theories which emphasize anterior processing (e.g., GNWT and HOT). However, the model shows that even without reporting demands, simulated PFC activity still reaches the threshold for reportability on some trials – maintaining the link between PFC activity and conscious access that is central to theories like GWT and HOT, and speaking to empirical finding from animal studies in which conscious contents can be decoded from PFC in no-report conditions (Kapoor et al., 2022).

The core idea implicit in these examples – and perhaps the kernel of a PP theory of consciousness *as such* – is that perceptual content is given by the brain's 'best guess' of the causes of its sensorium – this 'best guess' being the approximately optimal Bayesian posterior. The experiential character of this content is specified by the nature of the perceptual predictions in play. On this proto-theory, any change in conscious content must result from a change in the inferred state of the body, brain, and/or world (Hohwy and Seth, 2020, Seth, 2021, Whyte et al., 2022). (The reverse may not be true: changes in inferred state need not always lead to changes in conscious content.)

One appealing aspect of PP is its potential to go beyond highly constrained psychophysical environments to shed light on broader aspects of perceptual phenomenology that may be harder to account for by other theories. A good example is provided by perceptual (e.g., visual) hallucinations. The idea that hallucinations may result from overly strong perceptual priors dates back at least to Fletcher and Frith (2009) and was given computational precision by Suzuki et al. (2017) who used a neural network modeling approach to simulated the altered phenomenology that results from an imbalance between top-down predictions and bottom-up prediction errors. More recently, Suzuki et al., (2023) extended this approach to model the specific phenomenological characteristics distinguishing different kinds of visual hallucination – those arising from neurological disorders, visual loss, and psychedelics. By paying attention to more detailed aspects of phenomenology – beyond the simple detection and discrimination judgements widespread in most experimental settings – the PP approach can fulfill its promise to build explanatory bridges between (neural) mechanism and conscious phenomenology, with important clinical implications including but not limited to hallucinations.

Moving on, PP models have been extensively applied to aspects of conscious *self* – perhaps in contrast to other theories of consciousness (see Metzinger, 2004 for an early and influential attempt). The same strategy applies: the relevant conscious content is associated with an inferred state of the brain, body, or world – now with an emphasis on the body. An early example – albeit one that remains difficult to experimentally test – is the proposal that emotional experience corresponds to inference about the physiological condition of the body; a process of 'interoceptive inference' (Barrett and Simmons, 2015, Seth, 2013). In this view, the broad phenomenology of emotional experience – being dominated by valence – can be linked to a role for interoceptive inference in allostasis. This in turn connects closely to the FEP, according to which the entire edifice of perceptual inference can be derived from a fundamental imperative to continue existing – to stay alive. As I put it, "we perceive the world around us, and ourselves within it, with, through and because of our living bodies" (Seth, 2021). This perspective raises the intriguing but as yet untestable notion that consciousness may be a property of only (but not necessarily all) living systems – a form of biological naturalism (Searle, 2017, Seth, 2021, Seth, 2024b).

Grand claims aside, some progress has been made in computational models of specific aspects of self-related conscious experience. These range from early work by Stephan et al. (2016) linking disorders of allostasis to fatigue and depression to recent models of meta-awareness central to the flow of experience in contemplative states (Sandved-Smith et al., 2021). Rigorous testing of these and other models remains a challenge. As does extending PP to account for changes in conscious level, as in sleep and anesthesia. Here, it would be natural to appeal to the integrity of core mechanisms of prediction error minimization (Boly et al., 2011).

Whether PP succeeds as a theory of (or for) consciousness will depend both on evidence that prediction error minimization is indeed a core brain operation, and on its ability to draw explanatory and predictive links between elements of PP and properties of experience. While substantial evidence links top-down signaling to conscious perception (Hardstone et al., 2021), evidence for explicit sensory prediction error signals remains mixed (Solomon et al., 2021). And, while abundant evidence shows that participant expectations can shape conscious perception (de Lange et al., 2018) much remains to be done to causally connect the computational entities and dynamical processes of PP with specific forms of consciousness.

3. Where do we go from here: some insights about ToCs from the open discussion / Liad Mudrik & Lucia Melloni

Following the four talks (as mentioned, PP was unfortunately not presented at the meeting), an open discussion ensued. Below, we offer a selective overview of the key points of agreement and disagreements, open questions, and challenges to the theories that emerged from the debate. Overall, the discussion revealed some of the controversies and disagreements between the theories, while yielding few points of consensus.

3.1. What are we trying to explain?

One surprising point of agreement among the theories, considering previous discussions in the literature (Evers et al., 2024, Northoff and Lamme, 2020, Seth and Bayne, 2022), was their shared objective of explaining the phenomenology of the experience. That is, all theory proponents defined the explanandum of their theory as including both what distinguishes the presence vs. absence of an experience, and the phenomenal character of those experiences – why a certain experience feels the way it does, and how experiences differ. In that respect, an important clarification in the discussion was that both for GNWTand HOT, reports are not the explanandum itself, but rather a medium to capture the explanandum, which is phenomenology. Dehaene also urged the community to broaden its typically narrow focus on bottom-up visual perception and visual illusions, to other types of conscious experiences (e.g., feelings of knowing, awareness of making an error, etc.). It became apparent, however, that not all theories, in their present form, have provided explanations, or theory-based empirical investigations, of the phenomenal character of an experience. This holds true for HOT (though for a recent attempt, see Fleming and Shea, 2024, Lau et al., 2022), PP, but also to some extent to RPT and GNWT, where most empirical work has focused on the difference between consciously and unconsciously processed information.

The theories diverge however with respect to what that phenomenology is (e.g., detailed and rich but fleeting and non-captured by report, vs. sparse, low dimensional and fully captured by report) and what counts as phenomenological data (e.g., if an observation goes against folk-psychology, does it still count as data?).
Similarly, the theories strive to explain both *state consciousness* – what makes a person, or a system, conscious of its environment and its 'self' within that environment, as opposed to not being conscious – and *content consciousness* – what makes a person (or a system) conscious of a specific content at a specific time. This commitment to explaining the multiple aspects of consciousness is important, yet again, it does not fully align with how theories of consciousness have been studied thus far. As previously shown (Yaron et al., 2022), empirical support for the theories is highly unbalanced with respect to studies focusing on exploring state vs. content consciousness. Based on the updated ConTraSt database, which at present includes 503 experiments, but is mute with respect to PP due to lack of records (https://contrastdb.tau.ac.il), RPT has been almost exclusively studied with respect to content consciousness (97 %), and a similar bias also exists to some degree for GNWT 72 % of the experiments and HOT (71%), while IIT has been mostly studied with respect to state consciousness (81 % of experiments). Thus, for the theories to live up to their proclaimed explanandum, more diverse empirical research work is needed, focusing on the understudied domains for each theory. Similarly, for some theories, more theoretical work is needed, as acknowledged for both PP and HOTs above (notably, some claim that under specific frameworks of consciousness, the same explanation can account for both state and content consciousness; Aru et al., 2019; Bachmann and Hudetz, 2014. Yet these claims still require further empirical investigation).

3.2. What is consciousness?

Some of the differences between the theories appear to boil down to their different definitions of consciousness, which sometimes depart from common intuitions – or folk-psychological beliefs. For example, RPT explains away cases of inattentional and change blindness by claiming that participants actually did consciously see the unattended stimuli. That is, participants arguably had the conscious experience, although they don't know about it or don't remember it and so are unable to report it.

As Lamme himself acknowledged in the discussion, persuading the community that this counterintuitive interpretation is correct is a major challenge for the theory. Doing so depends on non-empirical arguments: if one finds the explanatory power of RPT high, they should also accept some of its non-intuitive claims, including this one. To him, for the field to stop going in circles, theories and their associated findings, particularly in neuroscience, must be able to trump some of our first-person perspective, folk psychological intuitions (Lamme, 2010), akin to the way researchers accepted non-intuitive claims in other fields, like quantum physics, when compelled to do so by empirical findings.

Yet this was far from being a consensual claim: Fleming argued that while quantum physics might reveal a counterintuitive truth about reality, a counterintuitive claim in consciousness science could take us away from the very same phenomenon we are trying to explain, which in turn might not even count as an explanation of the phenomenon. Thus, an open question is to what extent should theories of consciousness comply with folk-psychological concepts of consciousness. More specifically, the possibility of participants consciously experiencing unattended stimuli despite reporting otherwise was contested by Dehaene and Fleming, in line with their theories, while Boly argued that one might be conscious of some low-level features of the unattended stimulus, without necessarily experiencing an associated category.

Another example of the lack of agreement between the different theoretical stands relates to what happens when we are in a ‘flow’ state (Csikszentmihalyi et al., 2005) – for instance, reading ‘Crime and Punishment’, being immersed in challenging math exercises or even driving. What do we actually experience in these cases? And what differential predictions do the theories make about them (e.g., should there be a higher-order state for the experience of the environment while driving)? In the debate, we again failed to reach an agreement around these points, and it was suggested that there is no empirical way to solve some of these disagreements (e.g., about inaccessible experiences; Block, 2011).

Importantly, this discussion showcased that theories do not even agree on which states are conscious and which ones are not: for the very same experimental manipulation – inattentional blindness – RPT considers the information fully consciously perceived, IIT considers that some low-level features may be perceived while the category is not, while GNWT and HOT consider the information fully unconsciously processed (due to the lack of attention or higher-order representation, respectively). The divide went even further, as theorists were not able to agree on the criterion for detecting a conscious state: while for GNWT, some sort of reportability is required, for RPT the mere presence of perceptual organization suffices, and for PP the content needs to be encompassed in an inference about the state of the brain, body, or world. And, critically, we were unable to agree on a justification for either of these measures.

As Lamme put it during the debate, the current state of affairs seems more akin to differences in taste and preferences, where the very same data is held by one theory to reflect phenomenal experience, while another theory considers it to reflect unconscious processing. The lack of agreement between theories might render some of the claims almost tautological, such that they are always true, yet only within the framework of the theory. It is evident that this stalemate should be resolved for the field to move forward and to converge on a satisfactory explanation of consciousness. Yet no clear proposals were offered for how such resolution could be obtained.

On the one hand, this lack of agreement might be detrimental to the field: If the theories cannot even agree on the definition of what counts as a conscious event, it seems almost impossible to directly compare and test them. For that, the theories should be at least somewhat commensurable, both at the logical and at the empirical level (Evers et al., 2024). On the other hand, this might actually be a way to arbitrate between the theories: if one could present a conceptual, or an empirical, argument showing that inattentionally blind stimuli are either consciously or unconsciously processed, or that perceptual organization can (or cannot) take place unconsciously, this could serve as an argument against one or more theories.

More broadly, as a field, it might be worthwhile to explore consensus-establishing methodologies (e.g., Delphi studies; Barrett and Heale, 2020) aimed at agreeing on the definitive properties of consciousness and the associated data that could demonstrate their existence. Another constructive strategy might be to press theories of consciousness to provide clearer and more objective standards for what will count as evidence for or against their preferred definition. An alternative possibility is to accept that our understanding of the concept of consciousness is still within the prescientific stage. Accordingly, consciousness should be treated for now as a multidimensional entity, with different theories explaining different aspects of it. Then, more experimental and theoretical efforts will hopefully yield a more precise conceptualization. Under this approach, it would be advisable to systematically explore the parameter space of possible conscious experiences, and accept a more pluralistic view of this phenomenon (He, 2023). Such efforts might lead to the refinement and improvement of existing theories (Lakatos, 1978), or to a new account that would explain where existing ones are wrong. In parallel, bottom-up studies, collecting more evidence regarding the neural correlates of consciousness in a theory-neutral manner, could also be constructive in the attempt to establish a better, potentially novel, explanation for consciousness.

3.3. Are ToCs even theories?

According to Dehaene, the entire debate would not have taken place had ToCs been real theories, in the full-fledged sense of the term, as none of the current theories has achieved a level of theoretical development enabling precise formulations, leading to clear and explicit testable predictions that fully explain all aspects of consciousness. Dehaene suggested that, at present, perhaps a more modest term such as "framework" or "hypothesis" should be used – also for GNWT. According to him, the proliferation of theories stems from the partiality of each of them. What is lacking, he argued, is a precise formalization of the theories, leading to explicit simulations (as attempted for GNWT) that would allow us to test the theories in a much more rigorous manner. Consciousness research should accordingly move from descriptions to mechanisms, taking a much more mechanistic view. This could take place in two possible axes: at the anatomical level, we should strive for a more fine-grained description of the brain (e.g., at the synaptic or dendritic level; for a recent example, see Aru et al., 2020; Phillips et al., 2024). At the functional level, for theories that assume some form of computationalism, we should provide more detailed and accurate descriptions of the types of computations that are related to consciousness, specifying the relevant grain of the functional description, so that if they are implemented, consciousness should arise (e.g., Dehaene et al., 2017). If computationalism is not assumed, then theories should also be explicit about the non-computational functions associated with consciousness (Godfrey-Smith, 2016, Piccinini, 2020, Seth, 2024a).
Fleming considered the proliferation of theories and models a positive aspect of the field, rather than a disadvantage – according to him, it is a testament to the development of these accounts and their translatability into testable models, much like

in other fields (e.g., working memory). Yet he too acknowledged the limited nature of our theories; for HOTs, for example, the focus has mostly been so far on explaining presences vs. absence of stimuli, rather than accounting for the specific phenomenal aspects of experience (the same argument can be made towards GNWT and most theories of consciousness; Seth and Bayne, 2022). More recently, HOTs have been trying to develop that aspect of the theory as well, introducing the notion of a quality space that aims to model the relational properties of conscious experience (Lau et al., 2022, Rosenthal, 2010). Seth noted, retrospectively, that the (current) status of PP as a theory for consciousness science rather than a full-fledged theory of consciousness is an advantage here, mitigating against the temptation to overclaim for a theory while offering a continuity with an understanding of how brains and minds work in general. More generally, this discussion highlighted the lack of critical criteria for *what a theory must explain, to be counted as a theory of consciousness*; what does it have to provide to be regarded as a theory, as opposed to a description or a hypothesis (Doerig et al., 2020, Kuhn, 2024, Schurger and Graziano, 2022, Seth and Bayne, 2022)? In the same vein, it remains unclear what its explanandum/explanada should be. That is, should a theory of consciousness be required to explain *all aspects of consciousness* (to name a few: states of consciousness, contents of consciousness, the maintenance of a percept over time, the phenomenality of consciousness, its functions and its relations with other mental states)? Here, an interesting conflict arose between Lamme and Dehaene. The former claimed that GNWT or HOT do not explain perceptual organization, and hence cannot explain unity and integration, which he considers a key property of conscious experience. Conversely, Dehaene insisted, following Baars (1997), that this should not be an essential part of a theory of consciousness, which instead should only explain *conscious* perceptual organization (that is, it should not explain all aspects of the brain and its functions, but only the specific mechanisms that single out conscious vs. unconscious processing). Similarly, one can ask if a theory of consciousness must explain *all types* of conscious experiences, or whether it can be confined to one type, or one modality. And, perhaps most importantly, what level of prediction must it be able to generate, to be counted as a meaningful theory?

This question has been especially directed at IIT, in view of the claim that some of its aspects may be untestable (Barrett and Mediano, 2019, Doerig et al., 2019, IIT Concerned et al., in press). For example, it has been claimed that Phi (Φ) is not computable for large systems like the brain. To that, Boly replied in saying that for smaller systems, Φ is computable exactly (Albantakis et al., 2023), and for larger systems one can resort to approximations (she referred to a work in progress using fMRI focused at the voxel level - see Tononi et al., in press, though see Mediano et al., 2022, for a criticial discussion of approximations vs. proxies of IIT). She also referred to the PCI method as crudely capturing a proxy of Φ (though we note here that the results of PCI experiments are also compatible with other theories of consciousness, like GNWT; but see discussion in Tononi et al., in press). Finally, Boly argued that IIT makes many other testable predictions, about the factors leading to a loss of consciousness (Tononi et al., in press), or the substrate of specific kinds of experiences (such as space; https://osf.io/4rn85, and time; Comolatti et al., 2024), including counterintuitive predictions about the neural substrate of states of pure presence (Boly et al., 2024), and so on.

Seth later added that PP again offers an interesting contrast and alternative, providing a route towards a full theory of consciousness, rather than – as is presently the case – a set of resources for explaining properties of consciousness in terms of neural mechanisms and dynamics. One possible scenario for PP theories of consciousness is that the so-called hard problem of consciousness will not be directly solved (e.g., by a consensus agreement that process X generates or is identical to consciousness) but dissolved, as explanatory and predictive bridges between mechanism and phenomenology are built, tested, and refined (Searle, 2007, Seth, 2021).

3.4. What can refute the theories?

Perhaps the hardest question raised in this debate was "what would make you change your mind". Dehaene identified a central assumption of GNWT, according to which there should be only one central state of consciousness at a given moment. With this assumption, the theory tries to overcome a challenge facing local theories of consciousness – the problem of integration. Dehaene contrasted that with RPT, which focuses on the visual system but does not account for what happens within the other sensory cortices, or for non-sensory conscious contents (e.g. the sudden consciousness of having made an error). Arguably, RPT allows each sensory cortex to have its own feedback loops, potentially creating parallel conscious experiences for each modality. For GNWT, conscious experience must be integrated and central, as opposed to parallel. If one attends to audition, one cannot perceive a distinct competing visual stimulus at the same time, unless those two sensory inputs are integrated into a single unified percept (e.g. the McGurk illusion; McGurk and MacDonald, 1976). This is a key prediction of the theory, and if it is found to be wrong, the theory will be substantially challenged. According to Dehaene, this is a feasible test: experiments might show that GNWT underestimates the amount of parallel processing in the conscious brain, demonstrating that such processing can give rise to a conscious experience, as RPT claims. Or they could show that there is more than one central state or one central sharing system in the brain. All these outcomes would be highly informative for GNWT. We note that in this context, studies on agenesis of the corpus callosum (Paul et al., 2007) could be illuminating. Unlike the more classical studies on split brain patients (Gazzaniga, 1967), here the dissociation between the hemispheres is present from birth, providing a remarkable opportunity to investigate the possibility of multiple, concurrent, conscious experiences (using methods that do not rely on verbal report, given hemispheric disconnection). However, it still remains to be seen to what extent this pathological situation sheds light on the unity of consciousness in the neurologically intact brain.

Fleming maintained that we should not expect to find one critical piece of evidence that would refute HOT (or any other theory), and instead we should appeal to a plurality of measures and tests. According to him, there are many possible outcomes that would make him change his mind. For example, if it were possible to somehow inactivate or remove the neural substrate of the relevant higher-order representations, and participants would still claim awareness of the content that the higher-order state is pointing to, that would require a ‘reformulation’ of HOT, as he put it. Of course, this is currently not a technically feasible experiment, but according to Fleming, this might be achievable in the (near?) future. However, the real challenge with this proposal is that the theory currently does not specify an area that could be inactivated, as Fleming explained; since the higher-order state is a network property, it is less likely to be subserved by a specific region in PFC. This casts doubts on the feasibility of the proposed experiment, beyond the technical challenges. However, Fleming did mention that there might be a confidence ‘code’ that is expressed in localized activity patterns (e.g., Cortese et al., 2016; Masset et al., 2020) which one could then discover and potentially knock out. To the extent that subpersonal metacognition supports the relevant higher-order representations (e.g., Lau, 2019), this could provide a strong test of (a variant of) HOT, but so far this is still more of a theoretical conjecture than an empirical reality.

Boly specified several types of evidence that would challenge IIT: first, if one computed an approximation of Φ (based on the IIT 4.0 framework) and showed it is higher during generalized-clonic seizures, when brain activity increases but you lose consciousness, that would be a serious challenge for IIT. Similarly, if the approximations of Φ were found to be higher at the grain of molecules rather than neurons or microcolumns, or at the grain of microsecond or tens of seconds rather than the temporal grain of experience, that would seriously challenge the theory (Tononi et al., in press). The theory could also be indirectly challenged by results obtained through cruder proxies of integrated information, such as the perturbational complexity of EEG responses to TMS (PCI). PCI is meant to capture three of the five postulates of IIT: intrinsic existence (causation within the corticothalamic system), information (as estimated by differentiation), and integration (owing to the deterministic spread of perturbations within the corticothalamic system; Tononi et al., in press). Yet, as we further claim below, the PCI measure does not capture the full set of axioms of IIT and it is also compatible with other theories. Thus, in isolation it cannot be taken as either supporting or fully challenging IIT (but see again the discussion in Tononi et al., in press). Another line of investigation that could potentially refute IIT, according to Boly, focuses on the quality of consciousness: if the feeling of extendedness that characterizes spatial experiences is not subserved by grid-like networks in the cerebral cortex, that would be problematic for IIT (https://osf.io/4rn85).

Lamme reminded the discussants that he already changed his mind, as his original claim was that feedback must revert all the way back to primary visual cortex for consciousness to occur, something that he no longer holds as a necessary condition for consciousness. But the critical evidence that would make the theory false is twofold. First, finding recurrent interactions while the person is not conscious of the information (after excluding possible processes interfering with access or report to the conscious sensation, such as lack of attention, working memory etc.). To some extent, this is an almost trivial challenge, as it can be safely assumed that even in deeply unconscious states, some recurrent interactions between some neurons will remain. And also in subsystems that are generally considered not taking part in conscious experience, such as the spinal cord, recurrent interactions between neurons are present. Therefore, obviously, RPT will need some adjustments to better specify what the precise nature and extent of recurrent interactions have to be for conscious experience to arise. This has been referred to as the 'missing ingredient problem' in earlier publications (e.g., Lamme, 2018, where potential solutions have also been proposed). Second would be the finding of conscious experiences without recurrent interactions. If purely feedforward processes can support conscious experience, that would be a big challenge to the theory. This could be perhaps done using drugs or optogenetics, trying to selectively block feedback connections (Kirchberger et al., 2021). A similar challenge would arise if the signals that are typically considered markers of recurrent processing (such as delayed 'contextual modulation' signals, neuronal interaction measurements, etc.) are in fact due to feedforward processing. For example, delayed modulation of responses could be due to slow ascending subcortical arousal systems. Any mechanism where parallel streams operate at different speeds (such as magno- and parvocellular LGN inputs, for example) could mimic effects of feedback. It is therefore important to notice that the primary empirical observations on which RPT is based used contextual modulation signals for which a feedback origin has been confirmed using lesion experiments in monkeys (Lamme et al., 1998, Super and Lamme, 2007).

Finally, Seth argues that the utility of PP as a theory for, or eventually of, consciousness will turn on whether the core process of prediction error minimization can be experimentally verified or refuted. This is not trivial to do: the very flexibility of PP that endows it with the resources to address diverse aspects of conscious phenomenology also means that it is hard to identify any single (or set of) experimental tests that would be up to the job. For example, one might think that evidence against PP would be provided by experiments showing that conscious contents are constituted or deeply shaped by bottom-up signals, rather than by top-down signals. But the recent 'hybrid' extension of PP undermines this simple hypothesis, licensing instead a different set of predictions linking different types of phenomenology (e.g. focal vs gist) to different types (and directions) of predictive signaling (Tscshantz et al., 2023). Having said this, it should eventually be experimentally tractable to determine whether prediction error minimization is a core neural process. But this by itself is not enough. PP could also fail as a relevant theory if its resources – even if based on a sound empirical footing – do not, in fact, provide a flow of explanatory and predictive links between mechanism and phenomenology. In the terminology of Imre Lakatos (1978), PP will need to justify itself as *progressive* rather than *degenerate*, as a theory for, or of, consciousness.

Altogether, the proponents of these theories have offered potential experiments that could be carried out to refute their theories, given appropriate technical conditions. However, leaving technical feasibility aside for a moment, the major question is how these experiments are connected to the central ideas of the theories. The fact that the theories are still being developed and modified, and with the brain being the dumbfoundingly complex system that it is, theory proponents enjoy a high degree of freedom when interpreting data and making theoretical claims. For all theories, the bridging principles between the theoretical/computational claims (i.e., higher order thoughts, global broadcasting, integrated information, recurrent processing and predictive processing) and their biological realization are not fully specified. Thus, facing new data, the theories could be modified to accommodate these new findings, even when these do not confirm their predictions.

For GNWT, the functional neuroanatomy and physiology of the global neuronal workspace are still not fully clear and need to be further specified. How global and distributed should brain activity be to be considered as forming a conscious cell assembly within the global workspace? How many neurons should be activated and deactivated, and across how many areas? Which layers and cell types should be privileged? Since multiple superimposed active cell assemblies can co-exist in orthogonal prefrontal subspaces (Xie et al., 2022), what determines which one is conscious? To what extent should thalamic activity also be present? To what degree should these distant cortical and subcortical sites be synchronized, within which band (most likely beta and theta), and over which time period, in order to count as a conscious representation? Beyond existing findings on long-distance beta synchrony and causality (Gaillard et al., 2009), global information sharing (King et al., 2013) and vector stability (Schurger et al., 2015), this work on the formation of a conscious assembly should be brought down to the level of parallel single-cell measures, using both recording and causal stimulation methods.

For RPT, a similar ambiguity surrounds the scope of the required recurrency: how many neurons and areas should participate in a feedback loop for it to be conscious? Is V1-V2 feedback enough, or should the feedback encompass the whole visual system up to the

IT cortex? If so, how is the unity of consciousness achieved, considering that feedback will occur earlier and at different temporal intervals between closely connected areas and across the hierarchy? And how can integration be obtained between different contents, and between different modalities?

For HOT, it is still unclear how to map a theoretical distinction between first-order and higher-order representations onto anatomical predictions. Specifically, the theory does not provide explicit arguments for which properties of the brain should be regarded as the neural basis of consciousness. Thus, when facing a null result, it will always be possible to claim that some other brain area, or another type of processing, might still "host" the higher-order representation, yet it failed to be detected given the specific experimental conditions. More generally speaking, as explained by Fleming, the neuroscientific implementation of the theory is still being developed, so these questions might be answered by future formulations.

For IIT, the key challenge is that exact measurements of Φ cannot be obtained for large, multi-level systems. So far, tests of integrated information have relied on indirect measures, such as the PCI, which lack specificity because they are also compatible with other theories (for replies, see Tononi et al., in press). Approximations of Φ (based on IIT 4.0) that could be applied at the level of large-scale neuronal networks are being developed, though they have not been presented yet. Thus, currently, experimental attempts to test the theory (e.g., Cogitate Consortium et al., 2023; see also the INTERPID preregisteration: https://osf.io/4rn85) have accordingly focused on neural implementations of the theory, rather than on directly probing Φ. Other challenges are more conceptual: some criticized the axiomatic approach (Bayne, 2018), others claimed that the theory has consequences that they deemed improbable (e.g., that large grids could be highly conscious; Aaronson, 2014), to which IIT proponents replied (Tononi, 2014, see also the IIT discussion about grids above).

Finally, for PP, since the theory is far from being developed specifically for the purpose of explaining consciousness, it has yet to generate specific and clear predictions that can be directly tested. More generally speaking, if almost any circuit motif is compatible with PP and PE minimization, it is unclear what current finding can refute the theory. This theoretical flexibility leaves too much room for the theories to adjust in light of counterevidence. To make progress, we accordingly hold that the theories should provide much more detailed and explicit explanations of their core principles and how they translate into the more auxiliary predictions (i.e., how central they are to the theory; Chis-Ciure et al., 2024). Only with such specifications will arbitrating between theories be truly meaningful, pushing the theories towards substantial revisions – and even refutations – and not only minor ones. This process, which will surely take time, would allow the field to move from “surrogates for theories” (Gigerenzer, 1998) to fully-fledged theories that are well-specified and accurately defined.

3.5. Consciousness: matter/life or function?

A key challenge for the field in future years would be to face the question of what phenomenology actually is and which systems, biological or artificial, might have it. For computational functionalist theories (e.g., GNWT and variants of HOT), instantiating the right kind of computation(s) would be enough to instantiate consciousness (Dehaene et al., 2017; see also Butlin et al., 2023; Seth, 2024b). Dehaene offered an illuminating example, whereby under GNWT, a system such as a phone endowed with information sharing capabilities would instantiate consciousness. The defining factor, according to GNWT, is whether the apps in the phone are independent of each other (as they are right now) or have the ability to flexibly exchange information across apps in order to fulfill a certain goal (as a global workspace would permit). Quoting Dennett, Dehaene argued “*consciousness is a functional property that has all sorts of gradations and all sorts of adaptations*”. Yet, other theories (e.g., IIT) strongly oppose such computational/functionalist claims and instead offer structural explanations, according to which what matters for structure is how a system is built and not the functions that it performs, regardless of how sophisticated or smart those functions are (Findlay et al., 2024, Tononi and Raison, 2024, in press.). Specifically, IIT claims that computers that may replicate our behaviors or cognitive functions will not replicate our experiences (Findlay et al., 2024). Thus, according to IIT, a computer vision system in a self-driving car could act as if it did “see” functionally—recognize scenes and objects and move around the world much like we would—yet would not see anything phenomenally. Notably, computational functionalism (Butlin et al., 2023) is a stronger assumption than functionalism in general, such that non-computational forms of functionalism are possible (Piccinini, 2004). Some flavors of predictive processing, such as advocated by Seth, indeed challenge the assumption of computational functionalism and suggest that consciousness may depend on the material properties of living systems – a form of 'biological naturalism' (Searle, 2017, Seth, 2021, Seth, 2024b). According to some versions of biological naturalism, consciousness is a high-level emergent process that evolved in the complex, hierarchically organized living systems (Feinberg, 2024). How to test these highly divergent stands was left open and unanswered. When tackling this issue, a key question is what function – if any – does consciousness serve. Taking a broader, evolutionary, perspective in the years to come might help illuminate the question of the adaptative function of consciousness and its dependence or independence from biological systems (Cabral-Calderin et al., 2025, Cleeremans and Tallon-Baudry, 2022, Feinberg and Mallatt, 2016).

3.6. So, what are we left with?

A clear conclusion arising from the discussion is that there is currently significantly more disagreement than agreement between the theories. And the disagreement pertains to the most basic questions that define the scientific investigation to begin with: what consciousness is, which states are conscious and which are not, what is required from a theory of consciousness, and what exactly should be explained. As the theories themselves either do not provide full answers to these questions, or provide conflicting answers, it seems like it is up to the scientific community to establish a consensus on these basic questions, which might be a necessary condition for making progress in addressing them and in comparing and arbitrating among theories. Such consensus can be achieved in one of two ways; inter-theoretical dialogue between theory proponents, or – conversely – an initiative that stems from theory-neutral experts, who will lay out the grounds for future investigations and for a more informed evaluation of the theories. Once we agree on what we are trying to explain, and what is considered as an acceptable explanation, we can start weighing in on which explanations are better than others.
To do so, it would be useful to widen the scope of our investigation; thus far, experimental paradigms used to investigate consciousness have only probed some types of conscious experiences, with an overrepresentation of studies in the visual system (Yaron et al., 2022), mostly targeting perceptual consciousness using paradigms that are quite far from everyday conscious experiences (Mudrik et al., 2024). Thus, while much progress has been made in the field, we are just at the beginning. Much more empirical

work is needed both to characterize the minimal properties of consciousness and to understand how those properties evolve and what function they serve. Alongside theoretical development and refinement, it would be beneficial to invest empirical efforts in accruing more data, exploring the parameter space of the phenomena, and defining the boundary conditions of the observations.

Here, more than ever, patience might pay off. Developing coherent and systematic theories beyond frameworks and hypotheses takes time. Premature refutation might stifle theory development and progress. Like the taste of a good, mature wine, theories might also offer more satisfactory explanations if they are allowed time to mature. During the maturation process, these theories could make their assumptions, bridging principles across levels, and derivations clear, while embracing intellectual humility, given the high number of unknowns.

Landmark experiment sheds new light on the origins of consciousness

Date:

April 30, 2025

Source:

Allen Institute

Summary:

An unprecedented brain study has delivered fresh clues about consciousness—suggesting it's more about perception than planning. Two leading theories went head-to-head, but neither emerged victorious. Instead, scientists discovered that how we see may be more central to consciousness than how we think.

Share:

FULL STORY

A huge brain experiment just challenged what we thought we knew about consciousness—hinting it's rooted in how we perceive, not how we reason. Credit: Shutterstock

An experiment seven years in the making has uncovered new insights into the nature of consciousness and challenges two prominent, competing scientific theories: Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT). The findings were published today in *Nature* and mark a pivotal moment in the goal to understand the elusive origins consciousness.

IIT suggests that consciousness emerges when information inside a system (like the brain) is highly connected and unified, for as long as the information is consciously perceived, acting as a single whole. On the other hand, GNWT suggests a network of brain areas will spotlight important pieces of information in the brain -- bringing it to the forefront of our minds -- broadcasting it widely the moment it enters consciousness, and this produces conscious experience. The two competing theories were tested against one another in 2019 in a collaborative experiment involving 256 human subjects, and the findings were just released.

"Adversarial collaboration fits within the Allen Institute's mission of team science, open science and big science, in service of one of the biggest, and most long-standing, intellectual challenges of humanity: the Mind-Body Problem," said Christof Koch, Ph.D., meritorious investigator at the Allen Institute. "Unravelling this mystery is the passion of my entire life."

The Findings

Research showed that there's functional connection between neurons in early visual areas of the brain (the areas that process vision, which are at the back of the brain) and the frontal areas of the brain, helping us understand how our perceptions tie to our thoughts. The findings de-emphasize the importance of the prefrontal cortex in consciousness, suggesting that while it's important for reasoning and planning, consciousness itself may be linked with sensory processing and perception. In other words, intelligence is about doing while consciousness is about being.

This discovery has implications for how we understand consciousness and may shed light on disorders of consciousness such as comas or vegetative states. Identifying where consciousness comes from could help detect "covert consciousness" in unresponsive patients with severe injuries -- a condition known to occur in about one-quarter of cases as reported in the *New England Journal of Medicine* last year.

Neither Theory Came Out on Top

Integrated Information Theory (IIT) says consciousness comes from the interaction and cooperation of various parts of the brain as they work together to integrate information, like teamwork. It arises from how these parts are connected and how they share information with each other rather than any one individual area or part of the brain generating consciousness. The study, however, did not find enough lasting connections in the back of the brain to support this idea. Global Neuronal Workspace Theory (GNWT) supports the idea that consciousness happens in the front of the brain, but the study didn't find enough support for this idea either.

"It was clear that no single experiment would decisively refute either theory. The theories are just too different in their assumptions and explanatory goals, and the available experimental methods too coarse, to enable one theory to conclusively win out over another," said Anil Seth, Ph.D., a professor of cognitive and computational neuroscience at the University of Sussex. "Having said all this, the findings of the collaboration remain extremely valuable -- much has been learned about both theories and about where and when in the brain information about visual experience can be decoded from.

The study involved 256 subjects, which is unprecedented for this kind of experiment. Researchers showed them various visual stimuli and then used three common human brain measurement tools that track blood flow as well as magnetic and electrical activity to study their brains while they looked at the stimuli.

The highly collaborative experiment is the result of a large-scale, open science collaboration that began at a workshop at the Allen Institute in 2018. This innovative approach brought together researchers with differing perspectives to test two theories in a collaborative, yet critical, environment aimed at reducing confirmation bias and accelerating scientific progress.

"Adversarial collaborations are a powerful social process, little used because of its challenging nature, within any field that has competing theories," said Koch. "The bio-medical field could hugely profit by 'friendly' competition among theories -- neurobiological or others. But it requires a great deal of cooperation and work."

Materials provided by **Allen Institute**. Original written by Liz Dueweke. *Note: Content may be edited for style and length.*

Journal Reference:

1. Cogitate Consortium., Ferrante, O., Gorska-Klimowska, U. et al. **Adversarial testing of global neuronal workspace and integrated information theories of consciousness**. *Nature*, 2025 DOI: 10.1038/s41586-025-08888-1

Beyond the Familiar: Groundbreaking New Theories Shaking Up Consciousness Science

Discover the latest scientific inquiries and alternative models challenging our fundamental understanding of awareness.

The quest to understand consciousness, the very essence of subjective experience, is one of science's most profound and enduring challenges. For decades, theories like Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT) have been pivotal in guiding research. However, the landscape is rapidly evolving. Recent groundbreaking studies and novel theoretical frameworks are emerging, questioning the core tenets of these established models and paving the way for a new era of consciousness research. As of 2025, the field is abuzz with fresh perspectives that promise to deepen our comprehension of what it means to be aware.

Key Highlights: The Shifting Tides of Consciousness Research

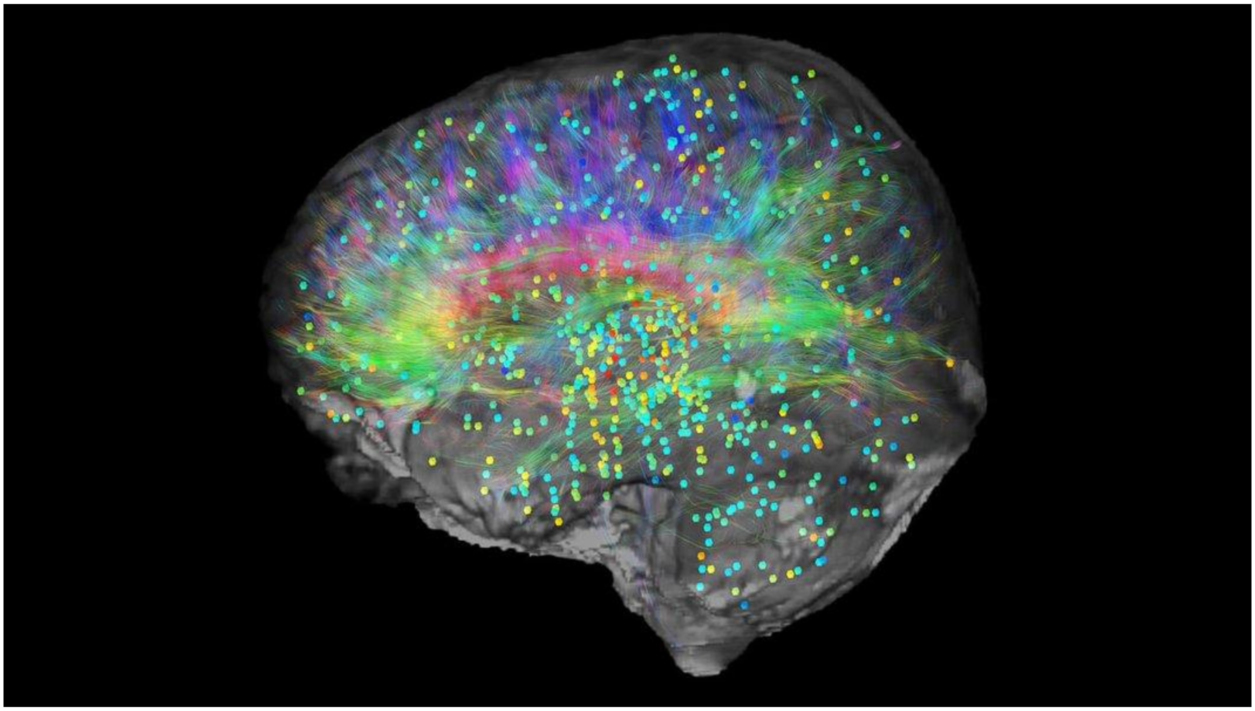
- **Landmark Experiment Challenges Giants:** A major seven-year adversarial collaboration, the Cogitate Consortium study, directly tested predictions from both IIT and GNWT, finding that neither theory's key predictions were fully supported by the experimental evidence, signaling a need for refinement or new approaches.
- **Emergence of Diverse Theories:** New models are gaining traction, ranging from theories grounding consciousness in basic cellular processes or embodied experiences to those proposing universal computational frameworks and reviving panpsychist ideas.

- **Focus on Broader Applicability and Testability:** There's a growing emphasis on developing theories of consciousness that are not only applicable to human brains but potentially to other biological organisms, artificial intelligence, and that offer more rigorously testable and falsifiable predictions.

The Cogitate Consortium: A Pivotal Moment for Consciousness Science

When Prevailing Theories Face Unprecedented Scrutiny

A monumental study, culminating in 2025 and involving over 40 scientists and 256 human subjects, marked a significant turning point. This seven-year project, orchestrated by the Cogitate Consortium and supported by initiatives like the Allen Institute, was designed as an "adversarial collaboration." Its primary goal was to pit two leading theories of consciousness, Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT), against each other by testing their distinct predictions using a battery of neuroimaging techniques (EEG, fMRI, MEG) while participants performed visual tasks requiring conscious perception.



A conceptual image representing complex brain activity, similar to that investigated in studies of consciousness.

Challenging Core Predictions

The results, published in prestigious journals, delivered a surprising outcome: neither theory's central predictions were unequivocally confirmed.

- **Integrated Information Theory (IIT):** IIT proposes that consciousness arises from the brain's capacity to integrate vast amounts of information, predicting sustained neural synchrony in posterior cortical regions during conscious experience. The study did not find consistent evidence for this specific type of sustained synchrony as predicted by IIT.
- **Global Neuronal Workspace Theory (GNWT):** GNWT posits that consciousness emerges when information is "broadcast" across a global network of brain regions, particularly involving

an "ignition" event in the prefrontal cortex. However, the predicted strong prefrontal cortex activity linked to conscious awareness, especially when stimuli entered or left consciousness, was largely absent or not as clear-cut as the theory suggested.

These findings don't necessarily invalidate IIT or GNWT entirely but suggest that they may be incomplete or require significant revisions. The study has invigorated the field, underscoring the complexity of consciousness and highlighting the urgent need for new theoretical frameworks or hybrid models that can better account for the empirical data.

Emerging Theories: Charting New Territories of the Mind

The limitations exposed in prevailing theories have created fertile ground for alternative and novel explanations of consciousness. These emerging frameworks often challenge fundamental assumptions about where consciousness resides, how it arises, and its very nature.

1. Cell-Based Theory of Consciousness

Consciousness Beyond the Brain?

A provocative new biological theory, gaining attention around 2023-2024, proposes that consciousness is not exclusive to complex nervous systems or brains. Instead, it posits that the origins of consciousness can be traced back to the very first cells, approximately 3.8 billion years ago. According to this view, fundamental cellular processes underlying awareness could mean that even simple life forms like bacteria, plants, and single-celled organisms possess a rudimentary form of consciousness.

Challenge to Prevailing Views: This radically expands the scope of consciousness, moving it from a phenomenon associated with higher cognitive functions in complex brains to a more basal and widespread biological property. It directly questions neurocentric views that tie consciousness strictly to neural architecture.

Implications: If substantiated, this theory could revolutionize biology and our understanding of life, suggesting that subjective experience, in some form, is an intrinsic part of being alive. It encourages interdisciplinary research bridging neuroscience, evolutionary biology, and cell biology.

2. Embodied and Bodily-Origin Consciousness Theories

"I Feel, Therefore I Am"

Challenging the traditional "I think, therefore I am" paradigm, several emerging theories emphasize the role of the body and sensory-motor processes in the genesis of consciousness. Research highlighted in 2025 suggests that consciousness might not primarily arise from higher cortical functions alone, but rather begins with fundamental bodily feelings, interoception (the sense of the internal state of the body), and interactions with the environment.

Challenge to Prevailing Views: These theories contest models that localize consciousness predominantly in specific brain regions like the frontal cortex or rely solely on complex neural network activity. They argue for a more distributed and integrated brain-body system as the seat of consciousness.

Implications: This perspective aligns with fields like neurophenomenology and enactivism, shifting the focus from isolated brain activity to the holistic experience of an embodied agent. It has implications for understanding emotional consciousness and the impact of bodily states on awareness.

The video above, "The Source of Consciousness - with Mark Solms," explores a theory that places emotions, rooted in bodily states and subcortical brain regions, at the center of mental life and consciousness. This perspective contributes to the broader discussion on embodied consciousness, challenging purely cortico-centric views by highlighting the foundational role of affective experiences originating from the body's interactions with the world.

3. Predictive Processing (PP)

The Brain as a Prediction Machine

Predictive Processing frameworks have gained significant traction in recent years. These theories propose that the brain is fundamentally a prediction engine, constantly generating models of the world and updating them based on sensory input. Consciousness, within this view, is closely linked to the processes of prediction, error signaling (when predictions don't match reality), and model updating.

Challenge to Prevailing Views: While PP can be seen as complementary to some aspects of GNWT (e.g., regarding information access), it offers a different emphasis. Instead of focusing solely on information integration (IIT) or global broadcast (GNWT), PP highlights the dynamic, hierarchical, and inferential nature of brain function as core to conscious experience. It suggests that what we perceive consciously is not a direct reflection of sensory input but rather the brain's "best guess" or active construction.

Implications: PP offers a powerful framework for explaining various perceptual phenomena, including illusions, hallucinations, and the subjective nature of experience. It's also being explored in the context of psychiatric conditions and artificial intelligence, as it provides a computational basis for understanding how systems make sense of their environment.

4. Conscious Turing Machine (CTM) Model and the Quest for Universality

A Universal Computational Theory?

Emerging from discussions around the need for theories of consciousness to be more universal and testable, the Conscious Turing Machine (CTM) model was proposed as a computational architecture. It aims to provide a framework for consciousness that is, in principle, implementation-independent, meaning it could apply to biological brains as well as potentially to artificial systems.

Challenge to Prevailing Views: This approach scrutinizes theories like IIT and GNWT against criteria such as falsifiability and universality (i.e., applicability to any dynamical system). While IIT claims universality, its practical computation for complex systems remains a challenge. The CTM, inspired by GNWT's global availability aspect but framed more abstractly, seeks to address these limitations.

Implications: Such computational models are crucial for advancing the study of artificial consciousness and for developing a more formal, mathematically rigorous understanding of conscious processing. They push the boundaries by asking what fundamental computational principles might underlie any conscious system.

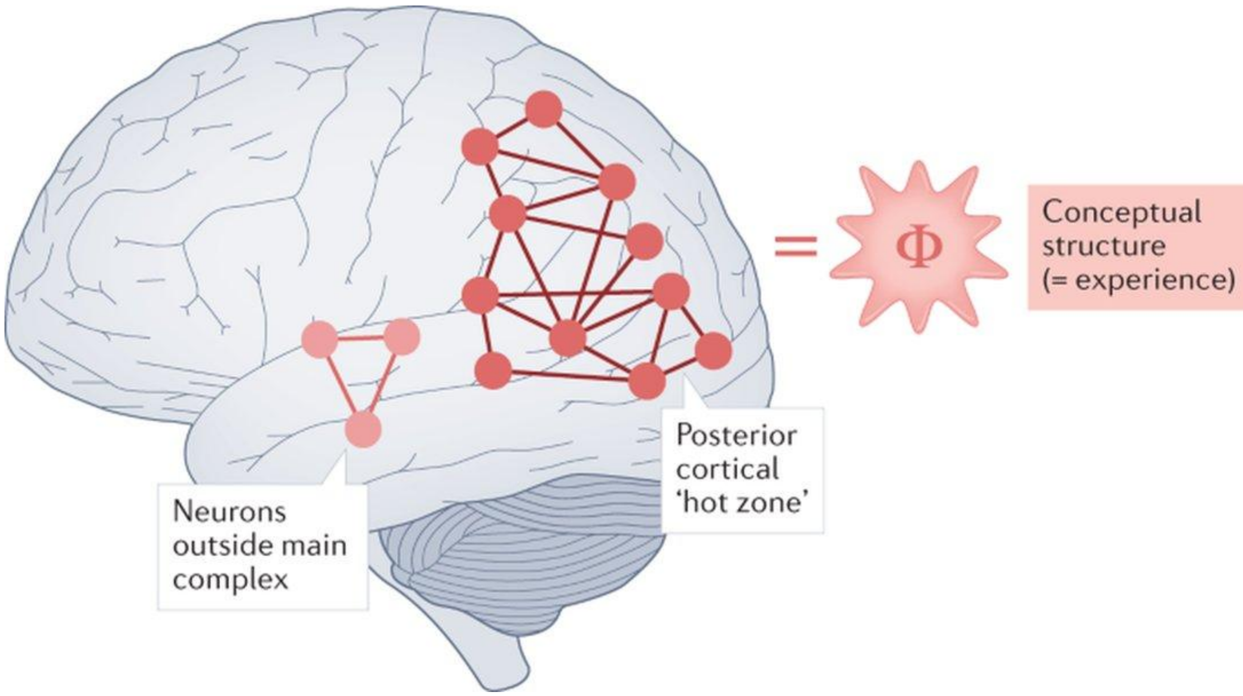
5. Panpsychism's Resurgence

Is Consciousness Fundamental?

Though an ancient philosophical idea, panpsychism has seen a resurgence in contemporary scientific and philosophical discussions. Panpsychism posits that consciousness (or some rudimentary form of it, often called "proto-consciousness") is a fundamental and ubiquitous property of matter, much like mass or charge. Complex conscious experiences, like human awareness, are thought to arise from combinations of these fundamental conscious entities.

Challenge to Prevailing Views: This directly challenges emergentist views (like IIT and GNWT, which see consciousness emerging from complex computations in non-conscious matter) by suggesting consciousness doesn't "emerge" from complexity but is an intrinsic feature of the universe at all levels.

Implications: Panpsychism attempts to address the "hard problem" of consciousness (why physical processes are accompanied by subjective experience) by dissolving it – if consciousness is fundamental, then its existence isn't something that needs to be explained as arising from the non-conscious. However, it faces its own challenges, such as the "combination problem" (how do micro-conscious entities combine to form macro-conscious experiences?).



A conceptual diagram illustrating various theoretical approaches to consciousness, highlighting the diversity in the field. (Adapted from general representations of consciousness theories).

Comparing Theories: A Radar Chart Perspective

To visualize how these different theories of consciousness stack up against various criteria, consider the following radar chart. This chart offers a speculative comparison based on general interpretations of these theories, focusing on aspects like their scope, testability, and how radically they challenge established thought. The values are illustrative and represent qualitative assessments rather than precise quantitative measures.

This chart illustrates how each theory might be perceived across different dimensions. For instance, "Cell-Based Consciousness" scores high on "Scope (Universality)" and "Challenge to Status Quo" due to its radical departure from brain-centric views, but lower on "Empirical Testability" with current methods. "Predictive Processing" shows strength in "Empirical Testability" and "Computational Clarity." Established theories like IIT and GNWT have moderate scores, reflecting their current standing and the recent challenges they face.

A Mindmap of Consciousness Theories

The following mindmap provides a visual overview of the relationships between prevailing theories, the pivotal challenges they face (such as the Cogitate Consortium's findings), and the array of emerging and alternative theories that are enriching the scientific discourse on consciousness.

Theories of ConsciousnessPrevailing TheoriesIntegrated Informatio nTheory (IIT)Consciousness as integratedinformationGlobal Neuron al WorkspaceTheory (GNWT)Consciousness asbroadcasted informa tionLandmark Challenges &DevelopmentsCogitate Consortium Stu dy(2025)Adversarial testing of IIT &GNWTIIT Predictions Not FullyCo nfirmede.g., specific posteriorsynchronyGNWT Predictions Not Full yConfirmede.g., prefrontal 'ignition'Emphasis on Universality &Falsif iabilityEmerging & AlternativeTheoriesCell-
Based ConsciousnessConsciousness in basic lifeformsEmbodied C onsciousnessBodily feelings asfoundationalPredictive Processing (P P)Brain as a predictionengineConscious Turing Machine(CTM)Univer sal computationalmodelsPanpsychism (Resurgence)Consciousnes s as afundamental propertyHigher-
Order Theories(HOT)Consciousness involvesmeta-

representationRecurrent ProcessingTheory (RPT)Importance of recu rrentneural activity

This mindmap illustrates how the field is branching out. The "Prevailing Theories" (IIT and GNWT) are being critically examined, particularly in light of "Landmark Challenges" like the Cogitate Consortium study. This has spurred the growth of "Emerging & Alternative Theories," each offering unique perspectives on the nature, origin, and mechanisms of consciousness, from the cellular level to abstract computational principles.

Comparative Overview of Select Theories

Understanding the nuances between these theories can be aided by a direct comparison of their core tenets and the challenges they face or pose. The table below summarizes key aspects of some prominent and emerging theories of consciousness.

Theory	Core Idea	Proposed Locus/Mechanism	Key Challenge to/from Prevailing Views
Integrated Information Theory (IIT)	Consciousness is identical to the amount of integrated information	Posterior cortical hot zone; specific patterns of neural	Challenged by Cogitate study on specific predictions of neural

Theory	Core Idea	Proposed Locus/Mechanism	Key Challenge to/from Prevailing Views
	generated by a system (Φ).	connectivity and dynamics.	synchrony. Complexity of measuring Φ.
Global Neuronal Workspace Theory (GNWT)	Consciousness arises when information is broadcasted to a global workspace, making it available to multiple cognitive processes.	Prefrontal cortex, parietal cortex, and long-range recurrent connections.	Challenged by Cogitate study on predicted prefrontal "ignition." Questions about sufficiency of broadcast for subjective experience.
Predictive Processing (PP)	Consciousness is related to the brain's continuous process of generating	Hierarchical cortical processing, interplay between top-down predictions and	Offers an alternative mechanism focusing on inference rather than

Theory	Core Idea	Proposed Locus/Mechanism	Key Challenge to/from Prevailing Views
	predictions about sensory input and updating them based on prediction errors.	bottom-up sensory signals.	just integration or broadcast; explains illusions well.
Cell-Based Consciousness	Rudimentary consciousness is a property of individual cells and fundamental life processes.	Basic cellular mechanisms of sensing, processing, and responding.	Radically challenges brain-centric views by extending consciousness to non-neural life.
Embodied Consciousness	Consciousness originates from and is deeply intertwined	Brain-body loops, subcortical affective systems,	Challenges cortico-centric and purely cognitive

Theory	Core Idea	Proposed Locus/Mechanism	Key Challenge to/from Prevailing Views
	with bodily feelings, sensorimotor interactions, and interoception.	somatosensory cortex.	views by emphasizing the body's foundational role.

This table highlights the diversity in current thinking. While IIT and GNWT have provided valuable frameworks, the newer theories are pushing the boundaries by considering different levels of biological organization (Cell-Based), different fundamental processes (PP), and the integral role of the entire organism (Embodied Consciousness).

-
-
-
-
-

On the Role of Theories in Consciousness Science

- Published: 26 November 2025
- Biyu J. He

Communications Psychology **volume 3**,
Article number: 173 (2025) Cite this article

Consciousness Science is entering an age of unprecedented opportunity, thanks to recent empirical and theoretical advances, increasing interest in the topic, and technological advances in neuroscience. The role theories will play in a maturing science of consciousness deserves a closer look.

"Theory without experiment is lame. Experiment without theory is blind."

— anonymous

Modern consciousness science was born ~35 years ago, when Francis Crick and Christof Koch helped usher in an age of empirical neuroscientific research on the "neural correlates of consciousness" around 1990[1]. Thirty-five years later, what role(s) should theories play in a maturing science of consciousness? This is a topic that has undergone considerate discussions, debate, and even controversy[2] recently.

Theories have always played a central role in consciousness science, owing, in part, to the field being situated at the intersection of philosophy, psychology, and neuroscience. The proposals of several prominent theories of consciousness—which remain influential today—including Global (Neuronal) Workspace Theory

(GWT/GNWT)[3], Integrated Information Theory (IIT)[4], and Recurrent Processing Theory[5] around 20–25 years ago helped build a strong community of researchers who conduct empirical or theoretical work on consciousness. Today, there are about five popular theories of consciousness, including the above three, along with higher-order theory[6] and predictive processing theory[7,8].

The nature of theories

What is a theory? Theory is defined as a plausible principle or body of principles to explain a phenomenon. A theory should be *simplifying* (otherwise it would not help with understanding) and have *predictive power* (otherwise it would be useless). A theory does not have to be comprehensive (e.g., a psychological theory does not have to specify neural implementation). A theory should have a clearly outlined explanatory scope (i.e., the target phenomenon it aims to explain). Finally, a theory should be testable in the sense that it can generate falsifiable hypotheses.

Theory is often confused with framework. Framework is a way to organize concepts, a way to think about a phenomenon. A framework tends to be broader than a theory and may or may not be directly testable. For instance, Marr's three levels of description (computation, algorithm, implementation) is a framework that has been extremely useful in cognitive science and neuroscience. A good example in consciousness science is the idea that sensory processing can be subliminal, preconscious, or conscious[9], which is a framework. By contrast, the idea that conscious access has the constituent mechanism of sensory information crossing a nonlinear bifurcation threshold (GNWT)[10] is a theory. Viewed in this light, IIT, at least in its original formulation[4], might be called a framework, because it is notoriously difficult to test but provides broader perspectives that can be quite helpful for thinking about consciousness and especially its enabling conditions (for instance, the idea that consciousness has both integration and differentiation—both a unified experience and a rich repertoire). When I proposed the Joint Determinant Theory (JDT) in 2023[11], I deliberately labeled it a framework because it offers a way to think about the neural basis of consciousness: a state-space view and the idea that the specific circuit mechanisms underlying different aspects of conscious content can be different; nevertheless, I retained "theory" in the name with the hope that over time it becomes specified with detailed neural mechanisms that are directly experimentally testable and the thinking that many contemporary "theories" are equally vague.

Compared to other disciplines in neuroscience and psychology, theories have played an outsized role in consciousness science. Empirical studies in this domain, and even grant proposals, are often expected to make contact with existing theories. Recently, $20–30 million was invested by Templeton World Charity Foundation in five adversarial collaboration projects carried out by large, international consortia to pit theories against each other. Sociologically, the few dominant theories have perhaps become entrenched, with young consciousness researchers often aligned with one of the camps. Pitfalls of this state of affairs have been pointed out recently, including confirmation bias that appears commonplace[12].

What is the best way forward? Some may advocate for empirical studies to ditch theories altogether. There are certainly great opportunities for theory-neutral empirical studies to contribute, especially at this (still young) stage of development of the field. But it is also important to remember that "experiment without theory is blind". With that in mind, here I advocate for a middle-of-the-road, nuanced approach.

The dangers of assuming uniformity

One potential key obstacle to progress is that consciousness theories are often idolized, their explanatory scopes not clearly outlined (some of the following ideas were recently discussed in ref. [11]). A common assumption is that a key principle, or set of principles, will explain all aspects of conscious awareness (the 'uniformity assumption'). Consciousness researchers often hold up the theory of evolution or the discovery of the double-helix structure of DNA as shining examples of the amount of explanatory power that a correct theory of consciousness will have. This is a romanticist vision, and it might be true, but it might not be; the jury is still out on this. The problem is that believing the uniformity assumption automatically extends the explanatory scope of any theory related to consciousness to encompassing all aspects of awareness, regardless of the origin of that theory. As a result, theories inspired by empirical observations of visual perception, sleep, metacognitive judgment, or attention are all automatically assumed to explain all aspects of conscious awareness. Should the uniformity assumption be wrong, this would cause massive confusion and gigantic roadblocks to progress, because theories that are most naturally suited to explaining different phenomena—

e.g., different aspects of conscious awareness—are pitted against each other in an endless cycle. This might also explain why which theory is supported by empirical results can be predicted by the experimental *paradigm* used by that empirical study[12], because if different theories have different explanatory targets that would be the natural outcome.

I believe that if we relax the uniformity assumption and allow theories to be humbler in their explanatory scope, the field would be able to make meaningful progress much faster and at the same time avoid a stagnant theoretical landscape. For instance, a theory most suited to explaining how a piece of sensory information accesses awareness does not have to simultaneously explain how sleep, psychedelic state, anesthesia and disorders of consciousness are all different from each other. A theory most suited to explaining conscious object perception (e.g., grouping) can be focused on explaining that particular phenomenon and other related phenomena, and be extremely useful at that. And a theory most suited to capturing the enabling conditions for consciousness to be possible in a system does not have to also explain how every aspect of awareness (e.g., perception and emotion) arises.

Such an attitude and culture would allow consciousness theories to be more focused and likely more experimentally testable. Put concretely, we need theories to explain how conscious perception occurs, how felt emotions happen, how conscious thinking unfolds, how intentions and agency are generated by the brain, and what brain state changes render consciousness impossible or altered under sleep, anesthesia, pathological and psychedelic conditions. This “boots-on-the-ground” approach also would allow new theories to be more readily proposed. Over time, when these "humbler" theories are experimentally validated, we would be on a firmer ground to examine whether there are common principles across these different aspects of consciousness. Importantly, such a shift in mindset would not diminish the significance of theories at all; it only takes away the grandiosity of theories.

Let us examine some examples of significant theories in other fields, which can help illuminate this discussion.

One of the all-time most influential theories in neuroscience is Hebbian plasticity proposed by Donald Hebb in 1949: "Neurons that fire together wire together." Since then, Nobel-prize winning experimental work has confirmed the existence of Hebbian plasticity in the brain. But empirical observations opposite to the predictions of Hebbian plasticity have also been found, in a phenomenon termed "anti-Hebbian plasticity"[13], which is thought to serve functions of preventing runaway excitation, decorrelating neural signals and maintaining excitation-inhibition balance. If the current model in consciousness science were adopted, the theory of Hebbian plasticity would have been "falsified" by these anti-Hebbian experimental observations and that would have been a huge mistake. Instead, Hebbian plasticity remains one of the cornerstone theories in neuroscience, with the appreciation that there are important exceptions to this rule.

Let us also take a look at an example with a much narrower scope: activity-silent working memory theory. This idea was proposed in a computational model in 2008[14]; since then it has become highly influential and has received direct experimental support[15]. This theory has a clear, focused explanatory scope: it proposes a specific neural implementation of a mental phenomenon (working memory), which *complements* rather than replaces an earlier theory on the same topic (persistent activity in working memory). I highlight this example here because the consciousness field needs more theories like this, which postulate concrete and testable neural mechanisms with clearly outlined explanatory targets. When theories are automatically assumed to explain all of consciousness, as is the case presently, they can become vague, shape-shifting, and difficult to test. Finally, from a sociological perspective, having humbler theories with clearly specified explanatory targets would also free the younger generation of consciousness researchers to come up with independent, fresh ideas instead of building their careers on testing existing dominant theories.

Consciousness science is at a time of unprecedented opportunities. No longer a tabooed topic and no longer marginalized, able to capitalize on the breakneck speed of technological advances in neuroscience, consciousness science is expected to flourish and deliver groundbreaking insights with enormous practical and clinical implications in the next few decades. This endeavor will be especially successful if we harvest the combined power of experiment and theory in a closed-loop manner: theories guide experiments and experiments lead to new or refined theories.

Front Sci, 29 October 2025
Volume 3 - 2025 | https://doi.org/10.3389/fsci.2025.1546279
This is part of an article hub

Consciousness science: where are we, where are we going, and what if we get there?

- 

  Axel Cleeremans [1,2*]

- 

  Liad Mudrik [2,3,4]

- 

  Anil K. Seth [2,5,6]

- 1. Center for Research in Cognition and Neuroscience (CRCN), ULB Neuroscience Institute (UNI), Université libre de Bruxelles, Brussels, Belgium
- 2. Brain, Mind, and Consciousness Program, Canadian Institute for Advanced Research (CIFAR), Toronto, ON, Canada

- See more

Abstract

Understanding the biophysical basis of consciousness remains a substantial challenge for 21st-century science. This endeavor is becoming even more pressing in light of accelerating progress in artificial intelligence and other technologies. In this article, we provide an overview of recent developments in the scientific study of consciousness and consider possible futures for the field. We highlight how several novel approaches may facilitate new breakthroughs, including increasing attention to theory development, adversarial collaborations, greater focus on the phenomenal character of conscious experiences, and the development and use of new methodologies and ecological experimental designs. Our emphasis is forward-looking: we explore what “success” in consciousness science may look like, with a focus on clinical, ethical, societal, and scientific implications. We conclude that progress in understanding consciousness will reshape how we see ourselves and our relationship to both artificial intelligence and the natural world, usher in new realms of intervention for modern medicine, and inform discussions around both nonhuman animal welfare and ethical concerns surrounding the beginning and end of human life.

Key points

- Understanding consciousness is one of the most substantial challenges of 21st-century science and is urgent due to advances in artificial intelligence (AI) and other technologies.

- Consciousness research is gradually transitioning from empirical identification of neural correlates of consciousness to encompass a variety of theories amenable to empirical testing.

- Future breakthroughs are likely to result from the following: increasing attention to the development of testable theories; adversarial and interdisciplinary collaborations; large-scale, multi-laboratory studies (alongside continued within-lab effort); new research methods (including computational neurophenomenology, novel ways to track the content of perception, and causal interventions); and naturalistic experimental designs (potentially using technologies such as extended reality or wearable brain imaging).

- Consciousness research may benefit from a stronger focus on the phenomenological, experiential aspects of conscious experiences.

- "Solving consciousness"—even partially—will have profound implications across science, medicine, animal welfare, law, and technology development, reshaping how we see ourselves and our relationships to both AI and the natural world.

- A key development would be a test for consciousness, allowing a determination or informed judgment about which systems/organisms—such as infants, patients, fetuses, animals, organoids, xenobots, and AI—are conscious.

Introduction

Understanding consciousness is one of the greatest scientific challenges of the 21st century, and potentially one of the most impactful for society. This challenge reflects many factors, including (i) the many philosophical puzzles involved in characterizing how conscious experiences relate to physical processes in brains and bodies; (ii) the empirical challenge of obtaining objective, reliable, and complete data about phenomena that appear to be intrinsically subjective and private; (iii) the conceptual/theoretical challenge of developing a theory of consciousness that is sufficiently precise and not only accounts for empirical data and clinical cases but is also sufficiently comprehensive to account for all functional and phenomenological properties of consciousness; and (iv) the epistemological and methodological challenges of developing valid tests for consciousness that can determine if a given organism/system is conscious. The potential impact of understanding consciousness stems from the many interlinked implications this can have for science, technology, medicine, law, and other critical aspects of society. Existentially, a complete scientific account of consciousness is likely to profoundly change our understanding of the position of humanity in the universe.

Accordingly, consciousness has become an object of intense scrutiny from different disciplines. While the connection between mind and body is an ancient philosophical conundrum, in recent decades, the metaphysical issues have been accompanied by a set of empirical questions, with neuroscience and psychology attempting to discover and explain the connections between conscious experiences and neural activity. Yet, strikingly, the core problem had already been formulated in scientific terms at the turn of the 20th century: certain articles from that period read almost as though they had been written today. For instance, in 1902, Minot wrote a *Science* article titled "The problem of consciousness in its biological aspects" in which he "[…] hopes to convince you that the time has come to take up consciousness as a strictly biological problem …" (1). Eighty-eight years later, Crick and Koch called for renewed inquiry into "the neural correlates of consciousness" (2, 3), prompted in part by the increasing availability of novel brain imaging methods that could link the biological activity of the brain with subjective experience. This empirical program continues apace, together with theory development and ever deeper interactions with philosophy. But today, there is also a sense that the field has reached an uneasy stasis. For example, a recent review (4) taking a highly inclusive approach identified over 200 distinct approaches to explaining consciousness, exhibiting a breathtaking diversity in metaphysical assumptions and explanatory strategies. In such a landscape, there is a danger that researchers talk past each other rather than to each other. Empirically, Yaron et al. (5) showed that most extant experimental research on theories of consciousness is geared toward supporting them rather than attempting to falsify or compare them, reflecting a confirmatory posture that hinders progress. This manifested both in the low percentage of experiments that ended up challenging theories, as opposed to supporting them (15%), and in the low percentage of experiments that were designed *a priori* to test theoretical predictions (35%, with only 7% testing more than one theory in the same experiment).

Beyond the genuine and highly complex scientific challenges that the study of consciousness must address, sociological factors may also contribute to the current sense of entrenchment: nobody likes to change their mind (6)! Emerging collaborative frameworks—especially *adversarial collaborations*—may help alleviate this concern, at least to some extent. But there are also further factors: the possibility that consciousness research is not sufficiently addressing *why* it feels like anything at all to be conscious and the role that conscious phenomenology plays in our mental, and indeed biological lives (7–9).

This paper is structured in a forward-looking manner, moving from the past, through the present, and on to the future. First, we clarify terms and make some essential conceptual distinctions. Then, we briefly review what has been achieved so far in elucidating the neural and theoretical basis of consciousness. Next, we consider the future of our field, outlining some promising directions, approaches, methods, and applications, and advocating for a renewed focus on the phenomenological/experiential aspects of consciousness. Finally, we imagine a time in which we have "solved consciousness" and explore some of the key consequences of such an understanding for science and society.

Three distinctions about consciousness

Consciousness is a broad construct—a "mongrel" concept (10)—used by different people to mean different things. In this paper, we stress three distinctions.

The first distinction is between the notion of the *level* of consciousness and the notion of the *contents* of consciousness. In the first sense, consciousness is a property associated with an entire organism (a creature) or system: one is conscious (for example, when in a normal state of wakefulness) or not (for example, when in deep dreamless sleep or a coma). There is an ongoing vibrant debate about whether one should think of levels of consciousness as degrees of consciousness or whether they are best characterized in terms of an array of dimensions (11) or as “global states” (12). In the second sense, consciousness is always consciousness *of* something: our subjective experience is always “contentful”—it is always about something, a property philosophers call intentionality (3, 13). Here, again, there is some debate over the terms, for example, whether there can be fully contentless global states of consciousness (14) and whether consciousness levels (or global states) and contents are fully separable (11, 15).

The second distinction is between perceptual awareness and self-awareness (note that in this article, we use the terms consciousness and awareness interchangeably). Perceptual awareness simply refers to the fact that when we are perceptually aware, we have a qualitative experience of the external world and of our bodies within it (though of course, some perceptual experiences can be entirely fictive, such as when dreaming, vividly imagining, or hallucinating). Importantly, mere *sensitivity* to sensory information is not sufficient to be considered as perceptual awareness: the carnivorous plant *Dionaea muscipula* and the camera on your phone are both sensitive to their environment, but we have little reason to think that either has perceptual experiences. Thus, mere sensitivity is not sufficient for perceptual awareness, as it does not necessarily feel like something to be sensitive. This experiential character is precisely what makes the corresponding sensation a conscious sensation (16).

We take self-awareness, on the other hand, to mean experiences of “being a self.” These experiences can be of many different kinds, from low-level experiences of mood and emotion (17) to high-level experiences of being the subject of our experiences, which might be supported by some inner (metacognitive) model of ourselves and our mental states (18–20). This kind of high-level reflective self-awareness is associated with the “I” and with a sense of personal identity over time (21).

The distinction between self-awareness and perceptual awareness is not sharp. Some aspects of the experience of “being a self” seem not to involve *reflective* self-awareness, such as experiences of emotion, mood, body ownership, agency, and of having a first-person perspective (22, 23). Some of these aspects may arguably have perceptual features. For example, emotional experience may depend on interoception (24–26). In addition, some perspectives, such as the higher-order theories described below, suggest that a form of metacognition might play a constitutive role in all instances of perceptual awareness, not only in self-awareness (18, 27, 28).

Human beings normally possess both perceptual awareness and self-awareness, but this is probably not true at all times or for all species. In humans, reflective self-awareness may be absent in specific conscious states, such as absorption or flow (29), or in states of minimal phenomenal experience (14). Other species may lack this reflective capability altogether. For example, few will doubt that dogs have perceptual experiences as well as various non-reflective self-related experiences—though this can be contested as we currently lack a way to directly test for consciousness in other species [see (30–32) for recent attempts to tackle this problem]. Nevertheless, there is no convincing evidence that dogs have reflective self-awareness in the sense defined above. Putting these debates aside, consciousness research has thus far largely focused, with exceptions (26, 33, 34), on trying to explain perceptual awareness as a first, albeit notoriously difficult, step toward understanding other aspects of consciousness. This emphasis most likely stems from the fact that perceptual awareness is generally easier to manipulate in experiments.

The third distinction contrasts the phenomenological (i.e., experiential) aspects of consciousness with its functions. This discussion has been largely shaped by Block's (35) influential, yet controversial (36, 37), distinction between phenomenal consciousness and access consciousness—informally, what consciousness *feels like* and what it *does.* Access consciousness is associated with the various functions that consciousness enables, such as global availability, verbal report, reasoning, and executive control. Phenomenal consciousness, on the other hand, refers to the felt qualities of conscious mental states: the complex mixture of bitterness and sweetness of a Negroni cocktail, the distinctive hue of International Klein Blue, the anxiety prompted by one's to-do list. All such conscious mental states have phenomenal character (using the philosophical term, often referred to as "qualia"): there is something it is like for us to be in each of these states. By contrast, there is nothing it was like for the neural network Alpha Go (38) to win against the South Korean world Go champion Lee Sedol (it was Sir Demis Hassabis and the DeepMind team who drank the champagne instead). Despite its seductive use of language, we think there is also nothing it is like for GPT-5 to engage in a conversation (39, 40).

Just as there has been greater emphasis within consciousness science on studying perceptual awareness compared with self-awareness, there has also been a greater emphasis on studying the functional rather than the phenomenological aspects of consciousness. This, again, may be due to the relative ease with which functional properties related to conscious access can be studied empirically compared with phenomenological aspects (41–43). With respect to the neural underpinnings of consciousness, we have been more focused on finding the mechanisms that differentiate between a consciously processed and an unconsciously processed stimulus than on explaining the difference between two conscious experiences, again with exceptions (44–48). Additionally, with respect to the functions of consciousness, we have been more oriented toward documenting what we can do *without* awareness rather than *because* of it (49–52). The potential for complex behavior in the absence of awareness has been further emphasized by the rapid advances in artificial intelligence (AI), where complicated functions can be executed without any accompanying phenomenology, at least as far as we can tell.

What have we achieved so far?

Following this clarification of terms, we briefly review where things stand today in consciousness research. Given the enormous challenge that explaining consciousness represents, it is easy to underestimate the significant progress that has already been made. This progress has been particularly visible over the last 30 or so years, but in fact it extends much further back, with highlights including seminal work on split-brain patients, neurological patients, work with brain stimulation, research on nonhuman primates, and much more (53–55).

Some basic facts are now well established. In humans and other mammals, the thalamocortical system is strongly involved in consciousness, whereas the cerebellum (despite having many more neurons) is not. Different regions of the cortex are associated with different aspects of conscious content, whether these are distinct perceptual modalities (56), experiences of volition or agency (34), emotions (57), or other aspects of the sense of "self" (58). Researchers have identified a myriad of candidate signatures of consciousness in humans, focusing on global neural patterns [e.g., neuronal complexity (59), non-linear cortical ignitions (60), stability of neural activity patterns (61)], specific electrophysiological markers of consciousness [e.g., the perceptual awareness negativity (62)], alpha suppression (63), late gamma bursts (64), and on relevant brain areas such as the "posterior hot zone" (65) or frontoparietal areas (66) as well as subcortical structures and brainstem arousal systems that may contribute to and modulate awareness (67–70). For some of these regions, notably brainstem arousal systems, there is debate about whether they represent necessary enabling conditions for consciousness and/or whether they contribute to the material basis of consciousness (67, 69).

At the same time, some previously popular hypotheses have now been empirically excluded. For example, the idea that consciousness is uniquely associated with 40 Hz (gamma band) oscillations has fallen out of favor based on substantial evidence (71, 72). In parallel, there has been a growing recognition that various confounds need to be carefully ruled out in order to interpret these findings, including those related to the enabling conditions for conscious experience, post-perceptual processes such as memory and report, and the concern that consciousness is often (but not always) correlated with greater signal strength and performance capacity (73–76). In this regard, phenomena such as blindsight, in which consciousness can be partly dissociated from performance capacity, are particularly intriguing [(77–79); but see (80, 81), for critiques].

Complementing these empirical findings, many *theories* of consciousness have been developed over recent years. These vary greatly in their aims and scope, in the degree of traction they have gained in the community, and in their level of empirical support (5, 12, 82–84). A selection of these theories provides a useful lens through which to focus attention on the progress made so far in the scientific study of consciousness.

Global workspace theory

One prominent theory, named "global workspace theory" (GWT), originated from "blackboard" architectures in computer science. Such architectures contain many specialized processing units that share and receive information from a common centralized resource—the "workspace." The first version of GWT (85) was a cognitive theory that assumed that consciousness depends on global availability: just like blackboard architectures, the cognitive system consists of a set of specialized modules capable of processing their inputs automatically and unconsciously, but they are all connected to a global workspace that can broadcast information throughout the entire system and make its contents available to a wide range of specialized cognitive processes such as attention, evaluation, memory, and verbal report (86). The core claim of GWT is thus that it is the wide accessibility and broadcast of information within the workspace that constitutes conscious (as opposed to unconscious) contents. Since the 1990s, GWT has developed into a neural theory (referred to as global neuronal workspace theory) in which neural signals that exceed a threshold cause “ignition” of recurrent interactions within a global workspace distributed across multiple cortical regions—this being the process of “broadcast” (64, 87). Importantly, GWT is what is called a first-order theory: what makes a mental state conscious depends on properties of that mental state (and its neural underpinnings) only and not on some other process relating to that mental state in some way. Thus, in contrast with the assumptions of higher-order theories (HOTs, introduced below), GWT does not postulate that consciousness depends on higher-order representation or indexing of some kind. GWT is primarily a theory of conscious access (88), focused on how mental states gain access to consciousness and how they accrue functional utility as a result. This is characterized largely in terms of supporting flexible, content-dependent behavior, including the ability to deliver subjective verbal reports [but see (89) for a discussion of the phenomenal aspect of consciousness and how the theory explains it, and see Dehaene’s section in (90)]. GWT’s clear neurophysiological predictions (centering on nonlinear “ignition” and on the involvement of frontoparietal regions) has led to a wealth of supportive experimental evidence (64). For example, divergences of activity ~250–300 ms post-stimulus have been associated with ignition (91), and measures of long-distance information sharing among cortical regions have been associated with broadcast (92). However, a major challenge for GWT lies in specifying what exactly counts as a “global workspace” (12): does it depend on the nature of the “consuming” systems, the type of broadcast, and/or on other factors?

Higher-order theories

A second prominent theory of consciousness is Rosenthal’s (93) higher-order thought theory, which proposes that a mental state is a conscious mental state when one has a “higher-order” thought that one is in that mental state. This core idea has now been elaborated on in different ways, resulting in a family of higher-order theories (HOTs). Unlike first-order theories, higher-order theories all claim that mental states are conscious when they are the target of a “higher-order” mental state of a specific kind (18, 93–95). The nature of the relationship between first-order and higher-order states varies among HOTs, but they all share the basic notion that for a first-order mental state X to be conscious, there must be a higher-order state X that in some way monitors or meta-represents X. Take the experience of consciously seeing a red chair. According to HOTs, the first-order representation (perhaps instantiated as a pattern of neural activity in the visual cortex) of red is not by itself sufficient to produce a conscious experience. Instead, there need to be additional “higher-order” states that point to or (meta)represent the first-order representation for it to be experienced as red. Crucially, such higher-order states need not be conscious themselves (i.e., we do not need to be aware of a mental state with content like, “I am now seeing red”). Rather, it is their very existence that makes the target content conscious. HOTs capture the intuitively plausible notion that a mental state is a conscious mental state as soon as I am aware of being in that mental state. This offers an equally intuitive distinction between conscious and unconscious mental states: I am conscious of some situation when I know about that situation; otherwise, I am unconscious of that situation.

Many HOTs locate the neural basis of the relevant meta-representations in anterior regions of the human brain, with an emphasis on the prefrontal cortex (96). Future "neural HOTs" will likely develop richer mappings between brain states and the theoretical distinction between first- and higher-order states (97). These theories are therefore supported by evidence implicating these regions in consciousness and undermined by evidence that anterior regions are not necessary for consciousness. As such, they have motivated studies investigating the neural correlates of consciousness (NCCs) with this question in mind (98). Of particular note are experiments that attempt to control for how well participants perform at a perceptual task: such studies (including in "blindsight" participants) have shown that when conditions are matched for performance, differences between conscious and unconscious perception are found in anterior cortical regions (75, 99) and interference with prefrontal function using transcranial magnetic stimulation (TMS) or multivariate neurofeedback affects subjective aspects of perception (such as confidence) without changing performance (100, 101). Studies associating perceptual metacognitive abilities with anterior prefrontal function also provide intriguing supportive evidence, albeit less direct (e.g., 102, 103). Additional support can be drawn from demonstrations of decoding of the content of consciousness from frontal areas (104).

However, HOTs currently do not fully specify the actual neural mechanism(s) mediating the implementation of first- versus higher-order states: how exactly does one brain state "point" at another, and what motivates the choice of which first-order state to point at or re-represent? Another challenge is that they focus on the contents of consciousness and provide less explanation for the level of consciousness. These under-specifications reflect the relatively limited empirical formulation of HOTs—despite their considerable philosophical backbone (105)—as compared with other theories (5). These aspects of the theory are currently being developed (45), and an ongoing adversarial collaboration (ETHoS[1]) is specifically aimed at comparing the empirical predictions of four HOT variants.

Integrated information theory

A very different perspective is provided by "integrated information theory" (IIT), developed by Giulio Tononi and colleagues since the 1990s (44, 106, 107). Rather than asking what in the brain gives rise to consciousness, IIT identifies features of conscious experience (described in five axioms) that it assumes are essential and then asks what properties a physical substrate of consciousness must have for these features to be present. A striking claim of IIT is that any physical substrate that possesses these properties will exhibit some level of consciousness (108). The two most illustrative essential features, or axioms, are (unsurprisingly) *information* and *integration*. According to IIT, every conscious experience is necessarily both informative (in virtue of ruling out many alternative experiences; i.e., every experience is the way it is, and not some other way) and integrated (every experience is a unified scene). IIT introduces a mathematical measure, phi (Φ), which, broadly speaking, measures the extent to which a physical system entails irreducible maxima of integrated information and thereby, according to the theory, provides a full measure of consciousness. Different versions of

IIT introduce different varieties of Φ, with the latest being IIT 4.0 (107), but all associate consciousness with the underlying "cause—effect structure" of a physical system and not just with the dynamics (e.g., neural activity) that the physical system supports. IIT is arguably the most ambitious theory we discuss because it addresses both the level and content of consciousness, proposes a sufficient basis for consciousness, and explicitly addresses phenomenological aspects of consciousness, such as spatiality (109) and temporality (110).

IIT has been criticized on the grounds that measurement of Φ is challenging or infeasible for anything other than very simple systems. Other "weak" versions of IIT have been proposed in which Φ is easier to measure, but this comes at the cost of abandoning claims of an identity relationship between Φ and consciousness (111). Another line of criticism is that the axioms proposed by full IIT do not satisfy standard philosophical criteria of being self-evidently true (112). Concerns like these have led to robust debate over whether the core claims of IIT are empirically testable and over what should be expected from a scientific theory of consciousness (40, 113, 114).

The most commonly referenced experimental support for IIT comes from evidence examining empirically applicable proxies[2] for integrated information (Φ) under different global states of consciousness. In a canonical series of studies (115, 116), Massimini and colleagues have developed a measure of consciousness, called the "perturbation complexity index" (PCI), which quantifies the complexity of the brain's response to cortical stimulation. Most commonly, the method uses TMS to inject a brief pulse of energy into the cortex, an electroencephalogram to measure the response, and the information-theoretic metric of Lempel–Ziv complexity (which quantifies the diversity of patterns within a signal) to quantify the complexity of the response. High PCI values arguably correspond to high levels of integration and information in the underlying dynamics. However, it is important to emphasize that the PCI, while inspired by and based on IIT, is not a *measure* or *approximation* of Φ, and differences in PCI across conscious levels may also be affected by differences in how unconscious processes operate at these levels. The PCI results, while fascinating, cannot be taken to directly support the distinctive aspects of IIT that rely on the definition of Φ, and are also compatible with or supportive of other theories, notably GWT. Nevertheless, the PCI method has shown exciting promise in important practical scenarios, such as detecting residual consciousness in unresponsive patients following severe brain injury (59).

In terms of neural correlates, IIT theorists claim that brain activity sufficient for conscious perception is localized to posterior regions (e.g., the posterior cortical "hot zone"). This claim is based on the argument that neural connectivity in these regions is well suited to generating high levels of (irreducible) integrated information, rather than the anterior regions favored by HOTs and GWT (117).

Predictive (and recurrent) processing theory

The final theory we mention here is not really (or at least not primarily) a theory of consciousness but rather a general theory of brain function—of perception, cognition, and action—from which more specific connections between brain processes and aspects of consciousness can be derived and tested (118). According to "predictive processing" (PP), the brain continually minimizes sensory "prediction error" signals, either by updating its predictions about the causes of sensory signals or by performing actions to bring about predicted or desired sensory inputs (the latter process being termed “active inference”) (119–121). This ongoing process of prediction error minimization provides a mechanism by which the view of perception as a process of Bayesian inference, or “best-guessing”, (122) and as a means of predictive regulation of physiological variables can be implemented (123, 124). In its most ambitious and all-encompassing version, the “free energy principle,” the mechanism of prediction error minimization, arises out of fundamental constraints regarding control and regulation that apply to all physical systems that maintain their organization over time in the face of external perturbations (125, 126).

Several distinct theories of consciousness fall under the umbrella of PP (e.g., 23, 127, 128). These typically share the claim that the contents of conscious experiences arise from (top-down) predictions rather than from a “read out” of (bottom-up) sensory signals. Informally, the contents of perceptual experience are given by the brain’s “best guess” of the causes of its sensorium or, even more informally, as a “controlled hallucination” in which the brain’s predictions are reined in by sensory signals arising from the world and the body (23).

One particular influential theory under the PP umbrella deserves mention: recurrent processing theory (RPT), also known as “local recurrency” or “re-entry” theory, associates consciousness with top-down (recurrent) signaling in the brain but does not appeal directly to the Bayesian aspects of PP (129, 130). Instead, RPT uses neurophysiological evidence to motivate the view that local recurrence (e.g., in visual cortex) is sufficient for phenomenal experience to occur and that feedforward (bottom-up) activity is always insufficient for conscious perception, no matter how “deep” into the brain this activity reaches (36). RPT’s focus on local recurrence is usually used to contrast the theory with other theories that involve widespread broadcast (GWT) or higher-order processes (HOT) (90), but as theories gain precision, it could be that aspects of RPT also surface in other theories (83). For example, the “ignition” process central to GWT might involve local recurrence. Nonetheless, a key difference between RPT and these other theories remains that RPT allows that phenomenal experience could be present without cognitive access (36).

The core commitments of PP do not directly specify a necessary or sufficient basis for consciousness to happen, nor do they specify how to distinguish conscious from unconscious processing. RPT is an exception here, proposing sufficient conditions, given the right enabling background conditions. Instead, the value of PP for theories of consciousness may largely reside in providing resources for developing and testing systematic or explanatory correlations between brain processes and properties of conscious experience, both functional and experiential (118). PP accounts tend to focus on conscious content rather than conscious level (e.g., 131, 132); they speak to both phenomenological (in terms of the nature of top-down predictions) and functional aspects of consciousness and address aspects of selfhood and embodiment more directly than other theories discussed here (e.g., 40, 133). Notably, variants of the theories discussed above can be expressed within the framework of PP, so there can be ‘PP versions’ of, for example, GWT and HOT (95, 134).

Whether PP succeeds as a theory in consciousness science will depend both on evidence that prediction error minimization is indeed a core brain operation and on its ability to draw explanatorily and predictively powerful links between elements of

predictive processing and aspects of conscious experience. While there is substantial evidence linking top-down signaling to conscious perception (135, 136), evidence for explicit sensory prediction error signals playing the roles proposed by PP remains mixed (137), at least when compared to the well-studied dopaminergic reward prediction error signal (138). Further, while abundant evidence shows that participant expectations can shape conscious perception (139), much remains to be done to causally connect the computational entities of PP with specific forms of consciousness. For some, this is a shortcoming of the theory: it might be too general and accordingly not informative enough to explain consciousness. Conversely, more specific formulations of the top-down principle, such as RPT, have been criticized for being too narrow, for example, focusing on visual processing only and failing to explain how this relates to other modalities and how conscious information is integrated across modalities.
This short tour of several of the many theories of consciousness [for a recent comprehensive survey, see (4)] highlights that there is not only a lack of agreement about the answers in consciousness science but also a lack of consensus about approaches and relevant questions. This does not mean there has been no progress. On the contrary, the last two decades have witnessed an enlightening move away from a simple search for NCCs in a comparatively theory-free and therefore explanatorily impoverished way to a rich landscape of different theories with varying degrees of experimental support. The Consciousness Theories Studies (ConTraSt) (**https://contrastdb.tau.ac.il**) database study has recently quantified the differences in the extent of research relating to the four theories of consciousness described above, and demonstrated how research results tend to align with the predictions of the supported theory [see **Figure 1** and (5)]. There are also some striking commonalities as well as differences among theories. For example, recurrent processing emerges as a key principle in GWT, IIT, PP, and some versions of HOT as well as other theories. Such unifying principles might point toward a “minimal unifying model” of consciousness, at least in biological systems (140).

Figure 1

Results of the Consciousness Theories Studies (ConTraSt) database study (5). Updated results of the ConTraSt database, now including 511 experiments published until mid-2025, which interpreted their findings in light of four prominent theories of consciousness: global workspace theory (GWT), higher-order theories (HOT), integrated information theory (IIT), and recurrent processing theory (RPT). Notably, there are currently no papers in the database for predictive processing theory (PPT). This is mainly because the database is based on the work done by Yaron et al. (5), where PPT was not included, and new uploads referring to this theory have not been made yet. **(A)** Distribution of experiments across theories. Green sections in the bars represent the number of experiments interpreted as supporting the theory; purple sections represent experiments interpreted as challenging it. **(B)** Effects over time: a cumulative distribution of experiments supporting the theories. **(C)** Functional magnetic resonance imaging (fMRI) findings for experiments supporting each of the theories. The same conventions used by Yaron et al. (5) are used here: for each activation, the color intensity indicates the relative frequency of experiments reporting activations in that brain area. While overlaying all findings demonstrates that most of the cortex has been implicated in consciousness, the breakdown by theory presents four different pictures, each aligning with the predictions of the supported theory. This further illustrates the confirmatory posture that most authors in the field have—intentionally or not—espoused.

Where are we going?

Thus far, we have surveyed some of the current main directions in the study of consciousness. As our overview makes clear, the sheer diversity of approaches and theories that characterize the field raises questions about how it can best make progress. In this section, we consider the most promising directions to follow in this ongoing quest, which some consider potentially endless (141). What will be the state of our field 50 years from now? Will our successors look back with satisfaction at the progress made toward "solving consciousness," or will they feel that the research has been going in circles, not getting any closer?

Considering that prophecy is given to fools, we will refrain from making a prediction here. But we note that the history of science abounds with unfulfilled scientific promises to solve one mystery or another, like producing cold fusion (142), curing cancer (143), achieving room temperature superconductivity, or indeed fully simulating the human brain (144). On the other hand, science often outperforms human predictions: 50 years ago, it probably seemed unthinkable that a computer would ever beat a human chess champion (145), converse fluently (146), or be able to create art (147). Bearing this in mind, what will the future of consciousness science look like? In the following sections, we sketch out nascent trends that will most likely shape the field in the coming decade: a shift toward theory-driven research, the necessity of collaborative and interdisciplinary work, the adoption of new methods, and an emphasis on applications. We hope that developments like these may help the field move beyond the current "uneasy stasis" we mentioned earlier.

From correlates to testable theories

The first major shift is a transition from "searching for the NCCs" to an increased focus on theory-driven empirical research (12, 82–84). While the former has been largely dominated by a data-driven, bottom-up approach consisting, for instance, of manipulating consciousness in hopes of identifying neural contrasts between consciously perceived and non-consciously perceived stimuli, the latter is driven by empirical predictions derived from specific theories of consciousness. Generally speaking, the agenda seems to be gradually transitioning toward providing explanations that go beyond descriptions [see (9), for a critical review]. This seems to be a step in the right direction, though more work is needed to potentially turn this simple step into a major leap.

First, theories must be thoroughly scrutinized to identify both their core constructs (148) plus testable predictions that have high explanatory power. Most, if not all, theories include claims and concepts that are somewhat fuzzy—often almost metaphorical—and these are then translated into neural terms in ways that are sometimes too simplistic, for example by debating whether consciousness is subserved by the front or the back of the brain (149, 150). Further elucidation and formalization are needed to make it possible for the theories to be fully tested. Addressing such issues would open up another research strategy, focused on the “search for computational correlates of consciousness” (151)—that is, identifying which computational differences best characterize the distinction between conscious and unconscious information processing. This in turn requires further precision. For example, what does it mean for information to be globally broadcast (152), and how do the receiving neurons understand the message? Similarly, how exactly does a higher-order brain state point at first-order brain states (96)? Or how is the unfolded cause–effect structure of a certain conscious state (107) physically implemented in neural terms? Only when predictions are fully fleshed out will we be able to assess their explanatory power using clear measures (153, 154).

Second, the explananda (explanatory targets) of the theories should be better defined, especially given claims that they might not be explaining the same things and the fact that they are supported by different types of empirical data, at least to some degree (5, 82). We believe that a greater focus on the phenomenological, experiential aspects of consciousness—for example, by studying quality spaces (45, 48, 155) or by pursuing computational phenomenology (14, 156–158)—is likely to yield substantial dividends here, by making the explananda more precise and thereby sharpening the distinctions among theories.

Third, as Seth and Bayne (12) argue, current theories should become not only more precise (for example, by using computational modeling) and more testable (for example, by developing new measures) but also more comprehensive. That is, theories should progressively be able to explain more distinct aspects of consciousness, and a good theory should explain as many aspects of consciousness as possible (82). An alternative and potentially complementary strategy is to focus on explaining the minimal, universally present features of consciousness (140)—perhaps reflecting a kind of “minimal phenomenal experience” (159).

Another shift in emphasis encouraged by theory-driven predictions is a focus on causal as well as on correlational evidence. Causal predictions generally provide stricter tests of a theory and hence more informative evidence. An example of a theory-based causal prediction can be found in INTREPID (**https://arc-intrepid.com/about/**), one of the current crop of adversarial collaborations. There, the team is using optogenetics in mice to contrast the effects of merely inactive versus optogenetically inactivated neurons in the visual cortex on visual perception, testing a prediction derived from IIT. Outside the context of theory testing, some have used optogenetics to examine the dependence of conscious perception of cortico-cortical and cortico-thalamic connectivity (157).

Finally, to allow us to home in on promising theories and reduce our credence in less useful ones, the field should focus on evaluating these theories through experiments designed *a priori* to test their predictions. At least some of these experiments should probe multiple theories simultaneously, to create meaningful contrasts between them. This leads us to the next suggested move.

From isolation to collaboration

Until recently, consciousness has mostly been studied by dozens of laboratories around the world, mostly independently. Each scholar has addressed the problem using their own tools, ideas, and theoretical approaches and pursued their research alone or with a small group. Yet, other fields have taught us that big questions often cannot be solved by individuals or small groups and that such questions may be better addressed through collaborative science (e.g., 160–162). Applied to our field, collaborative approaches can be used at multiple levels.

Selecting research questions

Defining key questions that are worth pursuing can be taken up by the community at large (163) or by a joint process involving multiple researchers and scholars. One form of such collaboration that we have already mentioned is *adversarial* collaboration, championed by Kahneman (6). Here, theoretical opponents work together to design experiments that would test their approaches, pushing each other toward better theoretical and experimental definitions of their claims.

A recent program initiated by the Templeton World Charity Foundation (TWCF) adopted this method in an attempt to "accelerate research on consciousness" by encouraging theory leaders to mutually engage and design experiments likely to arbitrate between competing theories. A series of such adversarial collaborations is now underway, pioneered by the Cogitate Consortium (**Figure 2**; 117). Time will tell if these collaborations allow us to arbitrate between theories. The first results of the Cogitate Consortium interestingly—and perhaps unsurprisingly—do not fully align with either of the predictions made by the theories in question, namely IIT and GWT. A challenge for this consortium, and likely for future adversarial collaborations, is that the agreed-upon experiments did not directly test the core aspects of either theory—a problem that in turn may follow from each theory making different assumptions and having distinct explananda. Yet, the experiments provided meaningful tests of the *neuroscientific predictions* of these theories, and the failure to confirm some of these predictions will hopefully lead to self-correction by the theories and to shifting the credence assigned by the community to each theory (164).

Figure 2

An illustration of the ongoing adversarial collaborations funded by the Templeton World Charity Foundation. Such collaborations invite theory leaders to jointly conceive experiments aimed at falsifying the core tenets of different theories. The experiment designs and theoretical predictions to be tested are preregistered and the experiments are performed and replicated by independent teams. In total, eight theories (see text) will be tested. To date, five adversarial collaborations have been launched. Cogitate (initiated in 2019) tested predictions of information integration theory (IIT) and global neuronal workspace theory (GWT). Data collection is complete and the first experimental results have been published (117). A second adversarial collaboration (2021) is comparing IIT and GWT in nonhuman animals. Thirdly, INTREPID (2022) is testing IIT against predictive processing theory (PPT) and neurorepresentationalism. A fourth collaboration (2020) contrasts higher-order theories (HOTs) of consciousness—specifically higher-order representation of a representation (HOROR) (94)—with some first-order theories, in particular recurrent processing theory (RPT) (130) and perceptual reality monitoring (PRM) (75). Finally, ETHoS (2023) aims to test four HOT variants: HOROR, PRM, higher-order state space (HOSS) (95), and the self-organizing metarepresentational account (SOMA) (18, 27). The outcomes of such vast empirical programs will likely shape the field over the next decade, but whether they will decisively rule out specific theories remains to be seen.

Defining research methods

Since its inception, the field of consciousness science has been characterized by controversies about how best to operationalize, manipulate, and measure consciousness for example, (165–168). Unsurprisingly, this lack of consensus practices is accompanied by a myriad of conflicting findings and claims, for example, on the scope of unconscious processing (169–171). Developing new methods and protocols that achieve broad uptake and consensus by virtue of having demonstrated validity would significantly advance the field (172), akin, for example, to collaborative attempts to define the goals of research in metacognition (173). Beyond making progress toward resolving key questions about what consciousness is and how it should be best studied, this would also allow direct comparisons between datasets obtained in different laboratories across the world and hopefully increase the chances of converging on agreed-upon claims about conscious versus unconscious processing. Notably, a single consensus approach would not suffice by itself and is likely to be extremely hard to obtain given the inherent complexities in studying consciousness. Rather, field-specific standardized approaches could usefully complement the rich variety of experimental and theoretical approaches currently flourishing. A relevant example is a recent collaborative effort (174) to define best practices for characterizing unconscious processing, for example, which awareness scale is preferable in each context, when tests of awareness should be administered, etc.

Collecting data

The plea for better-powered, multi-laboratory studies has been made in many fields, and in recent years such attempts have abounded (175–177). This might even be more crucial for consciousness research and psychological science more generally, where effects are typically weak and short-lived (178). Indeed, several such initiatives are already underway, some benefitting from the engagement and involvement of large swathes of the public. Examples include The Perception Census, a large-scale citizen science study of perceptual diversity (**https://perceptioncensus.dreamachine.world**), the SkuldNet COST Consortium (**http://skuldnet.org/**), and the Cogitate Consortium (117).

Interdisciplinarity

Consciousness is one of the most complex phenomena known to science, and understanding it requires the collaboration of scholars of different disciplines. In many ways, our field seems to have been a frontrunner in interdisciplinarity, as seen already in the “Towards a science of consciousness” conferences and at the annual meetings of the Association for the Scientific Study of Consciousness. Collaborations between neuroscientists, psychologists, and philosophers (11, 12, 48, 179–181); psychologists and computational neuroscientists (182); neuroscientists, philosophers, and physicists (183); and psychiatrists and psychologists (184) are just a few examples of the ways an interdisciplinary approach can advance the field of consciousness. Crucially, effective interdisciplinarity takes decades; this is perhaps one of the core arguments in support of bottom-up, curiosity-driven fundamental research: to allow interdisciplinary connections to formulate and flourish.

One aspect of interdisciplinarity worth highlighting is the benefit of involving philosophers of science (not only philosophers of mind). The challenge of understanding consciousness is of such magnitude that, even with substantial progress being made, there remain robust discussions about appropriate definitions, conceptual foundations, and constraints on empirical research as illustrated by the recent debate over IIT (40, 113, 114). Here, philosophy of science can provide a systematic meta-theory that can help the community to converge around exactly what should be explained and how.

Embracing new methods

Consciousness science could also benefit greatly from the development of new experimental methods, just as the field of neuroscience has benefited from functional brain imaging and other innovations. One promising arena for new methods is the opportunity to study consciousness in less constrained, more naturalistic environments. Given the challenge of studying consciousness, and a historical skepticism toward our ability to do so (185), the field has thus far mainly focused on finding the most controlled and simplified paradigms and operational definitions. Yet, recently, research has been conducted in more 'real world'-like settings, relying on state-of-the-art technologies that continue to evolve (186). Studies using virtual/augmented reality (187–190) suggest new ways to study consciousness, potentially in tandem with wearable brain imaging technologies such as optically pumped magnetometers (191). A particular advantage of these "extended reality" technologies is that they provide powerful new ways of investigating aspects of conscious experience that would otherwise remain difficult to study. Examples include experiences of embodiment through the use of avatars or virtual/augmented body parts (192–194), the influence of social context on the neural basis of consciousness, and the ability to suppress real-life objects (not merely computer screen images) from consciousness (195). In general, methods like these enable us to get closer to studying conscious (and unconscious processes) as they happen in real life, for example, when we are walking down a street—a situation of enormous sensory richness compared with a typical laboratory experiment (196).

Much can also be gained from combining the old and the new. An example here is the emerging approach of *computational neurophenomenology*, which merges new methods in computational models of neurocognitive systems together with relatively old philosophical and behavioral methods from phenomenological research to help build more informative bridges between brain mechanisms and conscious experience (151, 197). The key to doing so successfully may lie in flexibly recognizing how "old" and "new" approaches might benefit each other, rather than treating them in an either/or fashion, or as beholden to certain paradigms or explanatory targets.

One useful approach is to focus on the phenomenological aspects of consciousness. Here, a promising method to understanding *what an experience is like for an organism*—how it is similar or different to other possible experiences—is to investigate empirical mappings between (objective) neural and (subjective) perceptual similarity structures in high-dimensional spaces (45–48). As von Uexküll (198) and, more recently, Yong (199) point out, the world looks, smells, and tastes very different to a fly than it does to us: each organism is sensing its environment through sensory modalities that have been shaped by different evolutionary constraints and hence yield conscious experiences that are markedly different. Recent investigations using intracranial recordings have linked visual experiences to stable points in a high-dimensional neural state space (200–204). Such empirical investigations, in turn, bear on the philosophical claim that the qualitative nature of conscious experience arises by virtue of its relational similarity to other experiences—the “quality space” hypothesis (48, 205, 206). However, it remains unclear how these proposals for the neural encoding of qualitative features of (and relationships between) sensory experiences can be integrated into the theories of consciousness reviewed above [though see IIT for one approach (107)].

While these new methods offer exciting opportunities, they should not blind us to lessons from the past and lead us to forget hard-won lessons from earlier epochs of psychology (207). Of particular relevance to consciousness research is the issue of demand characteristics, which refers to how attributes of experimental context may implicitly influence participant behavior and experience (207). If experimental conditions are not carefully matched for participants’ expectations, observed differences may arise from mixtures of compliance and changes in experience caused by suggestibility and not from whatever other underlying process may be under examination (208, 209). Even more problematic may be experimenters’ own expectations about their study participants (210–212).

A focus on methods should also pay attention to the general problems of reliability and replicability that have plagued many areas of science, including psychology, where initiatives aiming to address these issues have become particularly prominent. Therefore, both old and new methods such as pre-registration, open data and materials, should be developed with appropriate rigor and embrace open science principles wherever possible. This probably means that most experiments will require large samples and more work such as measuring and controlling for demand characteristics, compared with what has typically been the case. This seems like a small price to pay to ensure, compared with what the field generates research of lasting value.

What if we succeed?

The final section of this paper ventures from the near future to the far horizon: imagine a time in which we have “solved consciousness,” whenever that may be. What would be the consequences of such an understanding? How would a complete explanation of the mechanisms of consciousness change the way we understand ourselves? One aspect seems clear—the more certainty we gain in our methods and in our theories, the more we will be able to translate findings from consciousness research into applications that address real-world problems. The impetus to apply consciousness science to address practical problems is increasing, given the accelerating progress in AI, developments in neuronal organoids (213, 214), and increasing societal attention to ethical concerns relating to nonhuman animal welfare and to the beginning and end of human life. For instance, a potential contribution of particular importance to society will be the development of a *test (or tests) for consciousness*, allowing one to determine, or at least provide an informed judgment about, which entities—infants, patients, fetuses, animals, lab-grown brain organoids, xenobots, AI—are conscious (32, 215, see also the section on “Clinical implications”).

Akin to calls to consider what may ensue if we succeed in building artificial general intelligence (216), here we consider the potential implications of success in consciousness science in four areas: scientific, clinical, ethical, and societal (**Figure 3**). Importantly, many of these implications may apply already even for partial “solutions” as we, hopefully, incrementally approach a full scientific understanding of consciousness.

Figure 3

Scientific, clinical, ethical, and societal implications of “solving” consciousness.

Scientific implications

Consciousness science remains somewhat marginal relative to the wider ecosystem of neuroscience and cognitive science (217). Every year, tens of thousands of neuroscientists attend the Society for Neuroscience Annual Meeting in the United States, but only a small fraction of their abstracts mention consciousness or subjective experience (in 2023, for instance, 92 abstracts included “consciousness”, compared with 4,297 that mentioned “behavior” and 7,237 that mentioned “brain”). Instead, the community’s approach to the brain remains mostly engineering-like: geared toward understanding how its component parts interact to produce behavior. This mechanistic approach will continue to accelerate in the next decades with the increased precision of tools such as optogenetics to probe and stimulate circuit function, affording unprecedented control over brain states and behavior. However, without progress in consciousness science, we may reach a point at which the brain, animal or human, is understood at a similar level as current AI systems. We might understand, at least in part, how it works, yet still not know how, or even whether, its mechanisms are linked to conscious experience. In contrast, success in consciousness science will provide a rich account of the biological basis of behavior and allow us to precisely determine when and whether certain brain states are conscious or underpin conscious experiences. In turn, it will provide neuroscience with the tools to control consciousness with the same precision as is currently being pursued for behavior[3].

Interfacing mechanistic neuroscience with consciousness science will be aided by asking what function(s) consciousness serves. Why should it matter that certain mental states are conscious and others are not if both are able to guide behavior? Indeed, much research on behavioral control and decision-making has proceeded without heeding consciousness as a variable (85). The key concepts of feeling, reward, value, valence, and utility have been approached differently in different fields, and have seldom been connected with consciousness research (33, 69, 218, 219). Debate continues about which broader behavioral functions specifically depend on phenomenal experience. Meanwhile, there is renewed interest in approaching this crucial “why” question from evolutionary (220, 221), philosophical (222), and psychological perspectives (8).

Clinical implications

Success in consciousness science will usher in a new realm of interventions in modern medicine. It already has a substantial practical impact on the clinical approach to neurological disorders of consciousness (223). Pioneering neuroimaging work over the past two decades has allowed communication with non-responsive patients previously considered to be in a vegetative state (224, 225). As discussed above, other approaches have applied indices of neural complexity, inspired by theories such as IIT, to distinguish between subgroups of patients with and without neural signatures that imply minimal levels of conscious experience (116, 226, 227). Thankfully, such cases of nonresponsive consciousness following coma are relatively rare in the population as a whole. In contrast, in an increasingly ageing population, the incidence of nonresponsive advanced dementias is likely to rise substantially. We know relatively little about what subjective experience is like in advanced dementia (228, 229), but progress in consciousness science will enable similar measures of consciousness level to be applied to such patients, to guide their care and to provide information to families about what their loved one's experience may be like.

Consciousness science also has considerable potential to improve our understanding and management of mental health conditions. These conditions, including depression, anxiety, schizophrenia, and autism spectrum disorders, are leading drivers of the global disease burden (230). In the European Union alone, the healthcare and socioeconomic costs of mental health conditions are estimated to total €600 billion/year (equivalent to >4% of gross domestic product) (231). Major unmet needs remain in this field, with first-line pharmacological and psychotherapeutic interventions, often discovered through serendipity alone, having remained unchanged for decades and showing limited effect sizes overall (232). One concern is that drug discovery efforts focus on behavioral markers of a certain condition (e.g., anxiety or chronic pain) in model animals such as mice, which may not accurately represent the conscious experience of these conditions in humans (233–235). Indeed, in 2017, Thomas Insel, a former head of the United States National Institute of Mental Health, conceded that even US$20 billion of investment had failed to "move the needle" on mental health (236).

From this perspective, it is striking that emotion and affective states in general have been comparatively neglected in consciousness research, (though see references 25, 33, 209, 237–241) and especially so if phenomenal experience is indeed about it *feeling* like something to be a conscious creature. Experimental medicine in this space is hampered by the fact that we still know very little about the neural mechanisms that underpin debilitating changes in subjective experience (242). Another interesting approach is to harness unconscious processes to develop more effective treatments for phobias. In one proof-of-concept study, researchers were able to decode fear-related representations that occurred unconsciously and reward them using neurofeedback (243). This, in turn, was reported to reduce physiological fear responses to consciously perceived stimuli without the participant having to undergo (often aversive) conscious exposure therapy.

The next frontier is to create an effective bridge from mechanisms to new interventions in mental health by studying markers of subjective experience in both human and animal models. Both sides of this interaction are crucial, as psychiatry and neuroscience need consciousness science and vice versa. Dysfunctions of consciousness in mental health disorders are probably distinctive in humans, having idiosyncratic content—mice may not become consciously depressed, for instance. At the same time, however, many of the same computational primitives supporting the neural control of behavior are conserved across species, raising the potential for biomarkers of conscious experience to be validated in humans and back-translated into animal models of mental health conditions (239–241). More broadly, if we gain a precise understanding of the biological basis of consciousness, it should be possible to design circuit-level interventions, for instance, brain–computer interfaces, that directly target, remediate, and potentially enhance aspects of conscious experience.

Ethical implications

Consciousness in nonhuman animals

Consideration of consciousness is often intertwined with ethical and moral obligations toward (presumed) conscious organisms (244). In ancient traditions, such as the Dhārmic religions, moral obligations toward nonhuman animals often rested on conscious aspects of experience (245, 246). For example, damage caused to an animal or system would only be ethically problematic if it caused them to be consciously in pain. Today, the corresponding philosophical term is "sentientism"—the notion that moral status follows from the capacity for phenomenal experience. Note that some versions of sentientism restrict the claim to the capacity for valenced experience, for example, the ability to feel some form of "good" or "bad" such as "pleasure" or "pain" [for a discussion of the importance of a general theory of suffering, see Lee (247) and Metzinger (248)]. Most consciousness researchers now reject the notion that only humans have consciousness (163)—a notion that can be traced in different ways to Aristotle and Descartes (249). However, they disagree on the dividing line between conscious and non-conscious entities, or indeed on whether such a dividing line exists at all. IIT, for instance, has implications that consciousness is probably a widespread feature of all living and potentially also some non-living complex systems (108). Conversely, global cognitive theories such as HOT suggest that a meta-representational neural system is a prerequisite for consciousness, one that might be limited in scope to those creatures who also have the capacity for, perhaps implicit, metacognition (75, 250, 251). As theoretical and experimental progress refines our credence in these various theories, we will be able to better characterize our confidence in which animals are conscious and what kinds of conscious experiences they may have (32). A consensual theory of consciousness would crystallize this dividing line, possibly affecting not only the use of model animals in neuroscience itself (244) but also societal perceptions of animals' suffering and their use by humans as sources of food, clothing, and medical products (252–255).

The ethical implications of success in consciousness science are intertwined with the kind of scientific explanation of consciousness that would correspond to such success. The example of life is informative. Vitalists held that there was a firm dividing line between the living and the non-living and that this was associated with some additional property, *élan vital*, that supported life. Biomedical science dissolved the need for such an extra property and led to a move away from "living versus non-living" being a central dividing line of moral status; for example, this can be seen in the adoption of brain death by Western medicine as the more relevant criterion for moral obligations toward sustaining life support. Indeed, consciousness itself now takes a more central stage in discussions of moral obligations toward humans and other animals. In the future, however, advances in consciousness science may begin to dissolve or reformulate those dividing lines based on consciousness, leading them to be replaced by other as yet unknown considerations.

Law

The law presents another broad area of societal implications. Many legal frameworks distinguish between notions of *mens rea* ("guilty mind") and *actus rea* ("guilty act"), in which *mens rea* picks out the conscious intent to engage in particular conduct. The neuroscience and biology of voluntary action, along with a deeper philosophical understanding of the concept of "free will", can be perceived as undermining the foundational notion that conscious intent is under the individual's control. Already, since the 1960s, brain injuries and diseases have been leveraged for legal defenses or *post-hoc* exoneration, as in the case of Charles Whitman, the "Texas Tower Sniper" found to have a large brain tumor impacting his amygdala (256). Today, the "my brain made me do it" defense is becoming increasingly popular, despite inherent conceptual difficulties (257). As we elucidate the neural basis of voluntary action (34, 258, 259) and the effects of unconscious processes on decision-making (260), it will become increasingly difficult to discern when moral and criminal responsibility should, if ever, apply (34, 261). These issues are not abstract. Whatever one's philosophical position on free will may be, judgments are being passed every day on people whose brain development and operation will have been affected by factors outside their control.

Artificial consciousness

Success in consciousness science may also result in a detailed understanding of mechanisms that can, in principle, be recreated in an artificial system (40, 262, 263). In philosophy, the position that mental states—including conscious ones—can be instantiated, rather than merely simulated, in an artificial system with different structural and/or material properties is known as substrate independence (or, in more restrained versions, substrate flexibility). Substrate independence/flexibility is closely related to multiple realizability—the idea that similar mental states can be implemented in different ways, although not necessarily in different kinds of material (264). Both substrate independence/flexibility and multiple realizability are, in turn, related to *functionalism*—the idea that mental states depend on the functional organization of a system, which can include its internal causal structure, rather than on its material properties (265). Within the category of functionalism, the more specific notion of *computational functionalism* claims that computation provides a sufficient basis for consciousness (266), so that consciousness could be implemented in non-biological information processing devices such as artificial neural networks of the sort deployed in cutting-edge AI systems (262). However, even if some version of computational functionalism is true, which remains debated (40, 267), abstracting the biological information processing sufficient for consciousness away from the messy realities of sensory and motor systems may be difficult, if not impossible, leading any artificial consciousness to end up looking animal-like rather than existing disembodied in software (268, 269). Further, functionalism, whether computational or otherwise, may not be true in the end, in which case other factors may be necessary for consciousness, such as being "biological" or "alive"—a position broadly known as *biological naturalism* (40, 270).

If artificial consciousness were achieved, whether by design or inadvertently, it would of course bring about a huge shift in allowing consciousness to decouple from biological life, which would in turn herald major ethical challenges on at least a similar scale to those discussed in relation to animals. The ethical problems could even be more severe in some regards since we humans might not be able to recognize or have any relevant intuitions about artificial consciousness or its qualitative character. There may also be the potential to mass-produce artificial consciousness systems perhaps with the click of a mouse, leading to the possibility (even if very low probability) of introducing vast quantities of new suffering into the world, potentially of a form we could not recognize. These observations provide some reasons for why we should not deliberately pursue the goal of creating artificial consciousness (271). There are other reasons too. Artificially conscious systems may be more likely to have their own interests, as distinct from the "interests" endowed by human designers. This could exacerbate the problem of ensuring their behavior is guaranteed to be in line with human and broader planetary interests: this being the "value alignment" problem already prominent in discussions of AI ethics.

In the near future, it is more likely that artificial intelligence, in the form of machine learning, and consciousness will continue to decouple. Such decoupling is likely to be counterintuitive and affect what we can conceive. For instance, the philosopher John Searle wrote in 2015, "[...] I am convinced that consciousness matters enormously. Try to imagine me writing this book unconsciously, for example." (272). Now, with the advent of large language models such as the GPT series, generative AI can write coherent and novel prose, and we can imagine how it could write a book unconsciously—though it might not yet be a very good one (273). Here, a different ethical question comes into focus. It is likely that artificial systems such as large language models will continue to improve in their mimicry of humans, leading a large sector of the population to adopt the "intentional stance" (274) toward such systems, attributing them with psychological properties such as beliefs, desires, and intentions, or perhaps even to fully perceive and believe these systems as being conscious. In these scenarios, people will intuitively assume they *are* conscious, even if scientists and engineers protest otherwise (40, 275–277). Such misfirings of mindreading machinery may sometimes be benign, for instance when children believe their favorite Pixar character is real (278–280), but this will not always be true. Significant challenges arise when people invest emotional significance into their relationships with seemingly conscious agents, as in the case of a Belgian man committing suicide after interacting with a chatbot (281).

More generally, the growth and societal penetrance of more powerful pseudo-conscious artefacts could have considerable societal impact (40, 282). People may be more open to psychological and behavioral manipulation if they believe the AI systems they are interacting with really feel and understand things. There may be calls to restructure our moral and legal systems around an intuition that an AI is conscious, even when it is not thereby potentially diverting resources and moral attention away from humans and animals that are actually conscious. Alternatively, we may learn to treat AI systems as if they are not conscious, even though we cannot help feeling that they are [perhaps illusions of artificial consciousness will be cognitively impenetrable, in the same way that some visual illusions are, (40)]. In situations like this, we risk brutalizing our minds, a danger long ago identified by Kant among others (283).

A key question running through all these issues is whether AI systems are more similar to humans in ways that turn out not to matter for consciousness (e.g., seducing our anthropic biases through linguistic ability) and less similar in ways that turn out to be critical (e.g., lacking the biological basis common to all known instances of consciousness, or not implementing the right kind of functional recurrency). A mature science of consciousness, guided by experiment and theory, will play a critical role in these debates.

Societal implications

An increasingly mechanistic understanding of consciousness is likely to reshape how humans see themselves and their place in the universe. We anticipate that human conceptions of consciousness will be reshaped by a gradual, mechanistic understanding provided by experimental science. The fields of evolution, genetics, and comparative cognition have eroded notions of mysterious human uniqueness in favor of explanations in terms of biological mechanisms. We envisage a continuation of this process as consciousness research continues to mature (though of course, a backlash is always possible, if elements of society feel threatened). This form of "success" that we envisage in consciousness science may lead the moral and ethical implications of consciousness to become more nuanced. In 1950, Alan Turing suggested that "[…] at the end of the 20th century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (284). Perhaps a similar process will lead us to soon no longer be vexed by the question, "But is it conscious"?

More problematic may be the societal and ethical questions that accompany the ability to *control* consciousness in other animals and humans (22), which is similar to how the development of clustered regularly interspaced short palindromic repeats (CRISPR) technology has allowed biomedical scientists unprecedented artificial control over life. There, the concern is not whether the RNA strands or bacteria going into a CRISPR-cas9 system are categorized as living or non-living. Instead, the overriding concern is that humanity now has the power to determine who and what lives, and the parameters, such as genetics, that affect the trajectory of such a life (see also 285 for a perspective based on bioelectrics rather than genetics). The molecular biology community has led the way in creating globally agreed standards on which forms of genome editing are permissible and which are prohibited—standards that have largely been adhered to, though with some exceptions. By analogy to CRISPR, similar concerns may be raised by a model of consciousness that is detailed enough to allow lasting, systematic manipulations of subjective experience—for instance, through pharmaceutical or brain–machine interface stimulation (286). Here, too, there will be a need for robust frameworks for scientific governance surrounding what types of “consciousness editing” are permissible, whether in the manipulation of consciousness where it already exists or in the *de novo* creation of conscious systems (248, 263). There will be related ethical questions about what forms of consciousness are ethically desirable or undesirable. These issues may arise sooner than we think. The rise of brain organoids and assembloids as model systems for neuroscience already pose a challenge in understanding how and whether synthetic consciousness may be created in a lab (214, 287, 288).

Conclusions

From Copernicus to Darwin, from Freud to Turing and modern biology and neuroscience, the history of scientific endeavor has repeatedly dethroned humanity from the center of the universe, each time widening our wonder rather than reducing it, and each time rendering our view of ourselves as more *part* of nature rather than *apart* from nature. There is every reason to believe that a deeper understanding of consciousness will follow a similar trajectory, enriching our lives with meaning and beauty, rather than draining them of these things.

But every revolution in understanding is different, and consciousness will be different again. How will a complete scientific understanding of consciousness penetrate our everyday conscious experiences and our appreciation of the human condition? As *we ourselves* are the explanatory target for consciousness science, might it be difficult, perhaps impossible, for us to fully appreciate the explanatory power of a successful theory of consciousness, even if one is developed? In this scenario, a distinctive cognitive disconnect may persist for consciousness science, so that our scientific understanding remains divorced from our lived experience. The alternative is that our experience of being human will change in ways that are currently extremely hard to foresee. One possibility is that we come to see ourselves as more fully embodied, rather than as conscious minds carried about by meat machines (289).

Closer to home, a mature and adequately funded (290) science of consciousness will intervene in many contemporary debates, perhaps resolving them decisively or at least changing their nature. Discussions over how to treat animals, adult human patients, and unborn children will be substantially informed by knowledge of the degree and form of conscious experiences in these organisms. The current politically dominated discourse in these areas may come to be seen as archaic as religious debates about the nature of the solar system seem to us today, but this, of course, depends on how political attitudes evolve.

Just as humans are now beginning to be able to create life from scratch, we will also be able to create conscious minds from scratch. What will this new ability to “play God” do to us? Importantly, we may be able to create specific kinds of consciousness rather than just new conscious organisms as we do when having children.

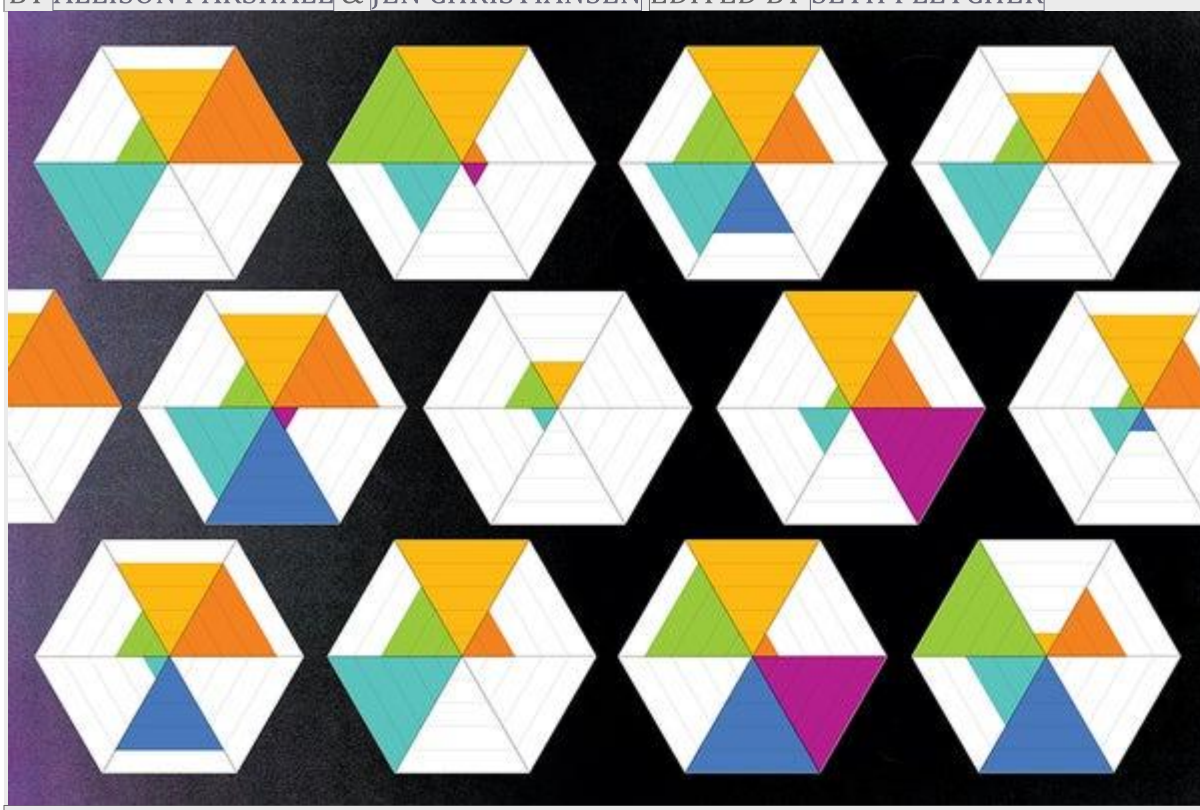
Finally, it is tempting to wonder what other discoveries could follow or precede the scientific explanation of consciousness—discoveries that could match its potential for reframing the human condition. Predictions here are especially futile, given the prevalence of unknown unknowns, but one possibility is finding evidence of extraterrestrial intelligent life. Such a discovery could highlight the diversity of conscious minds, the uniqueness of our own, and change how we see ourselves within the vastness of the universe. The difference between a universe teeming with mere life and one suffused with awareness is simply astronomical.

January 20, 2026
2 min read

Your guide to 29 wildly different theories of consciousness

The many, many ways researchers hope to solve the toughest mystery in science

BY ALLISON PARSHALL & JEN CHRISTIANSEN EDITED BY SETH FLETCHER



Jen Christiansen

February 2026 Issue
Consciousness

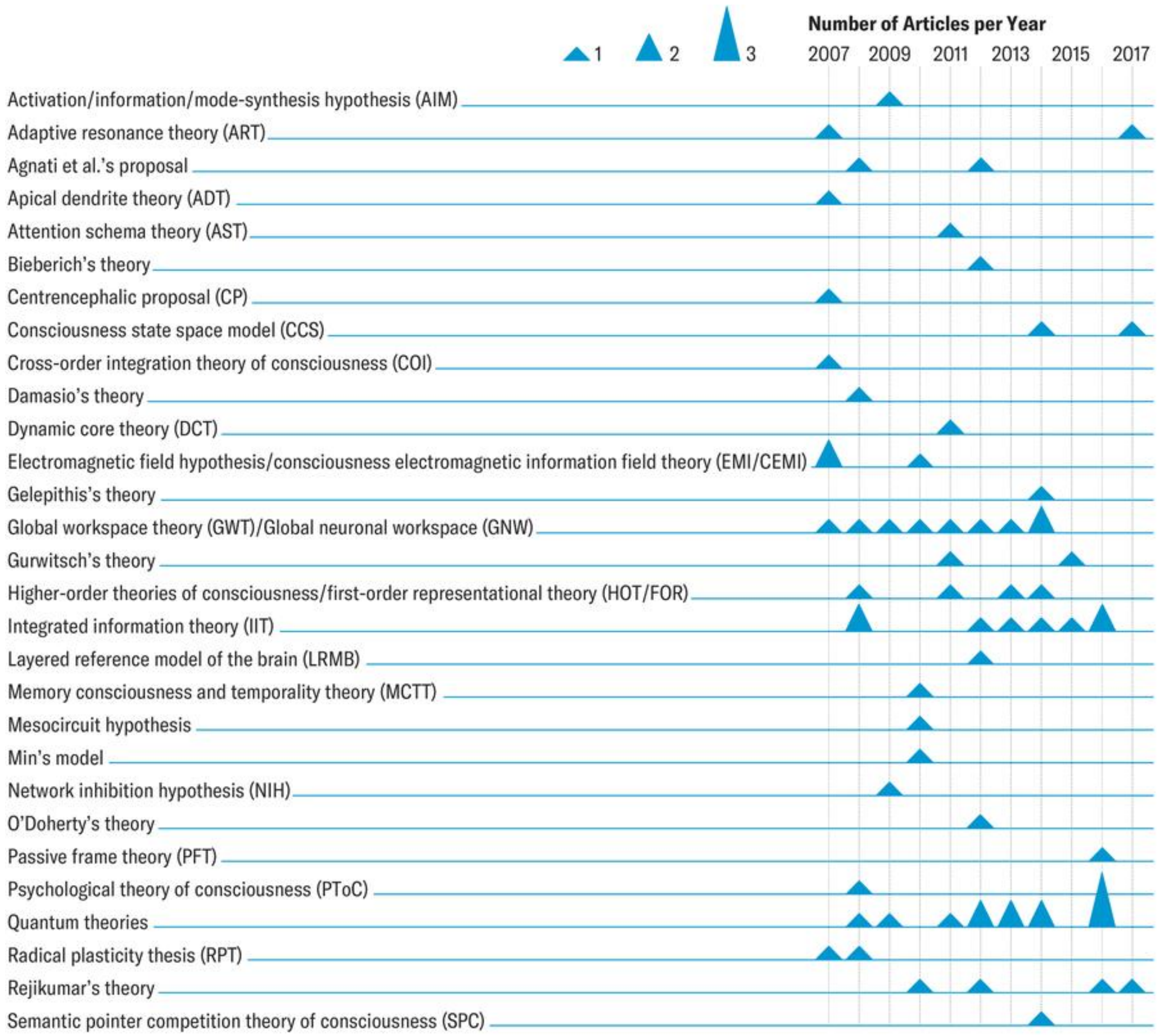
This graphic is part of a special report on consciousness. Read the full feature here.

Theorists have proposed many explanations for consciousness. They are startlingly diverse, with different goals, starting points and even definitions of consciousness. To compare them, researchers surveyed articles suggesting a theoretical model for consciousness that were published between 2007 and 2017 in English or Italian.

Some of these theories have many publications to their name, such as global workspace theories, higher-order theories and integrated information theory, three of the field's leading models. But having more publications does not mean a theory is supported by stronger evidence—this analysis didn't weigh the strength of study results. Instead the number of articles most likely reflects the amount of interest each type of theory garnered. This may explain why quantum theories of consciousness, which are fascinating but not yet grounded in much evidence, have been proposed so many times. Predictive processing theories, which are also influential in the field, are notably absent, perhaps because they originated as theories of perception rather than consciousness.

THEORIES INCLUDED IN THE META-ANALYSIS

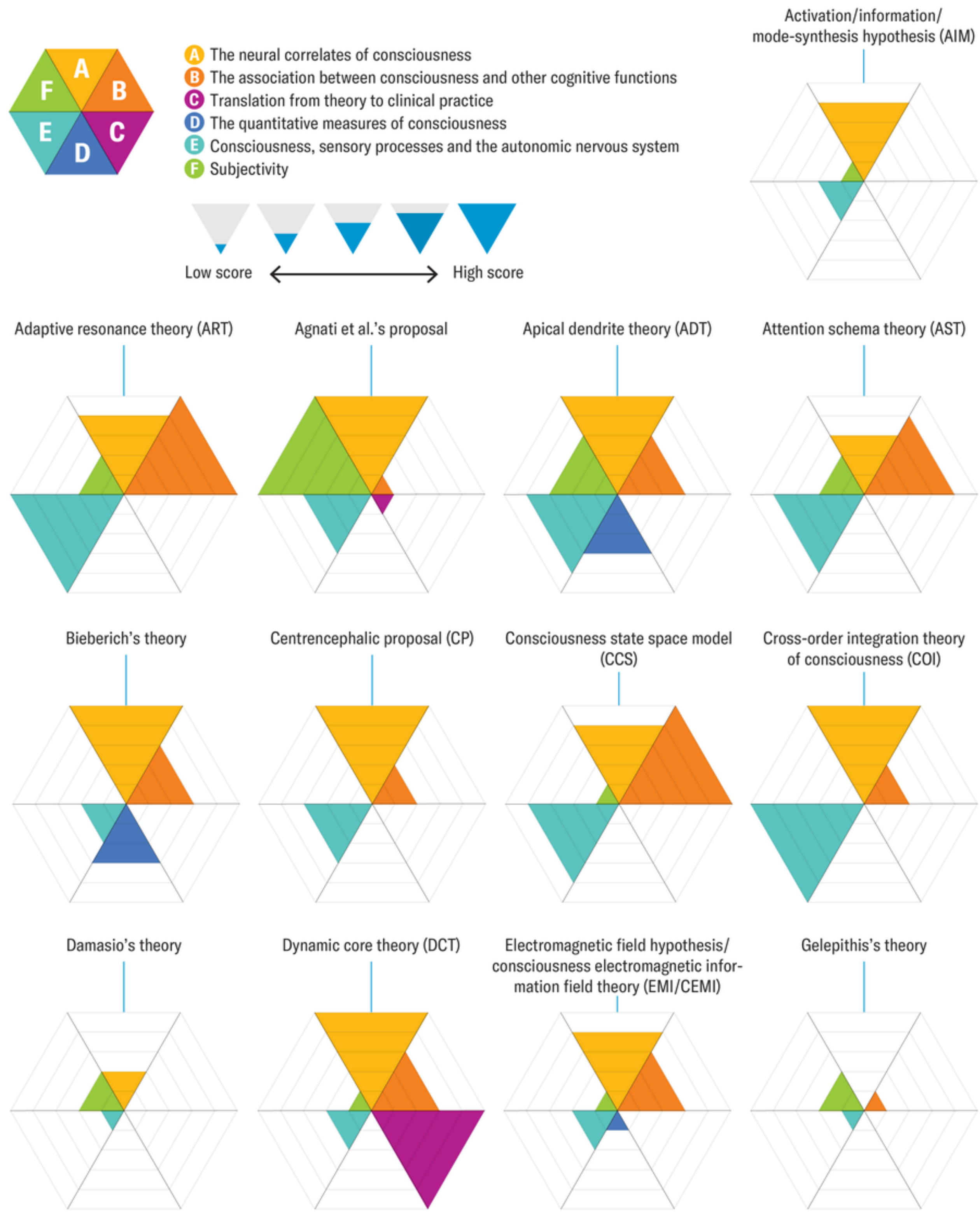
The researchers found 68 articles that analyzed theoretical aspects of consciousness and that were published between 2007 and 2017 in English or Italian. Together, they represent 29 different types of theories. More than half (18) were suggested only once, but others garnered more interest.

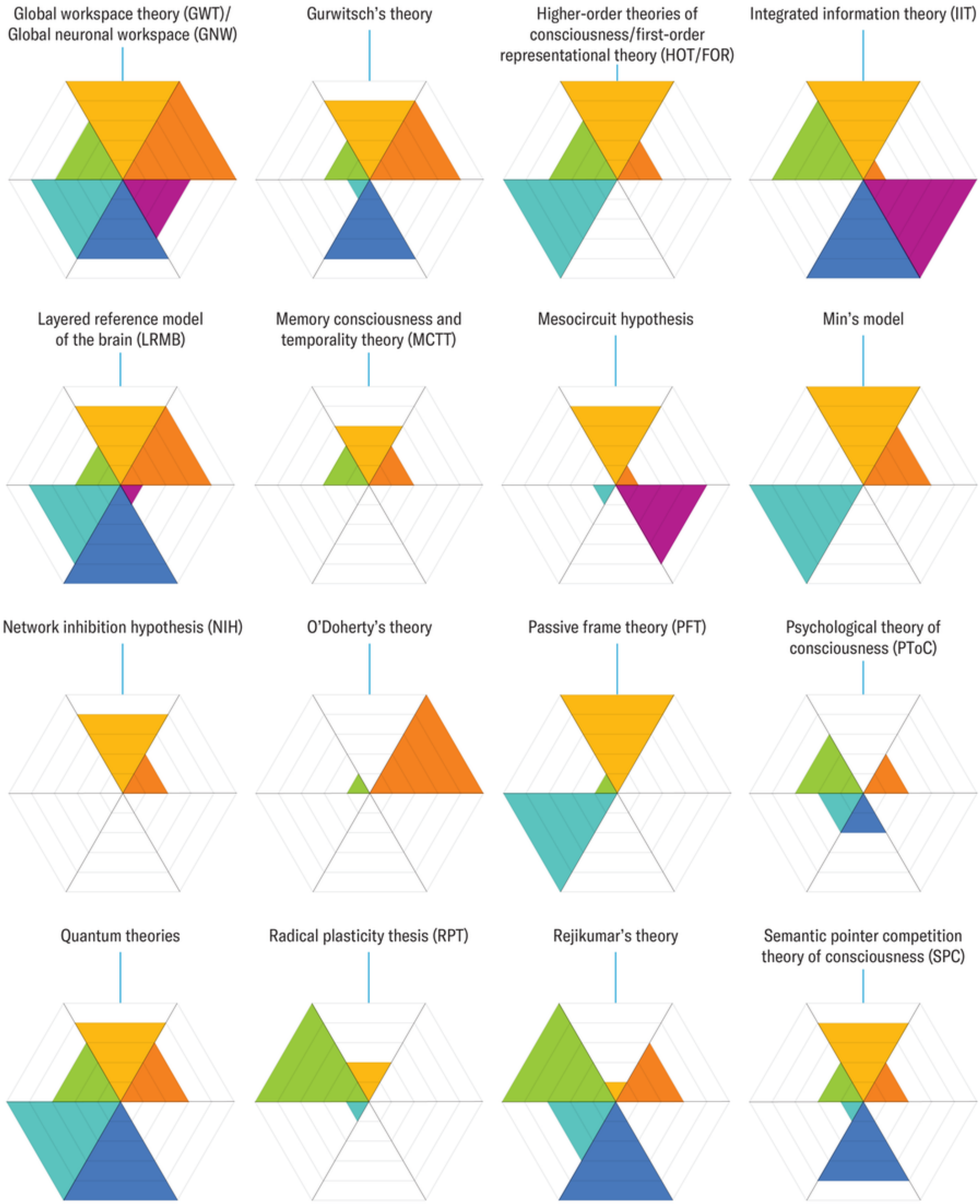


Jen Christiansen; Source: "Theoretical Models of Consciousness: A Scoping Review," by Davide Sattin et al., in *Brain Sciences*, Vol. 11; April 2021 (*data*)

THEORY CHARACTERISTICS

All of these proposals are wildly different, and only some are supported by evidence. The researchers analyzed how each tried to address six dimensions of consciousness: (A) what brain activity correlates with conscious experience, (B) how consciousness relates to other mental faculties such as memory and attention, (C) the clinical treatment of disorders involving consciousness, (D) how levels of consciousness could be measured, (E) how consciousness relates to sensory information, and (F) where subjective experience comes from.




Global workspace theory (GWT)/ Global neuronal workspace (GNW)
Gurwitsch's theory
Higher-order theories of consciousness/first-order representational theory (HOT/FOR)
Integrated information theory (IIT)
Layered reference model of the brain (LRMB)
Memory consciousness and temporality theory (MCTT)
Mesocircuit hypothesis
Min's model
Network inhibition hypothesis (NIH)
O'Doherty's theory
Passive frame theory (PFT)
Psychological theory of consciousness (PToC)
Quantum theories
Radical plasticity thesis (RPT)
Rejikumar's theory
Semantic pointer competition theory of consciousness (SPC)