Integrated Predictive Workspace Theory: Towards a Unified Framework for Consciousness Science

Rui Lin
Lin.Rui.ipwt@proton.me
Independent Researcher

June 26, 2025

Abstract

Consciousness science seeks to understand how the brain unifies information into conscious experience. Existing theories (IIT, GWT, PCT/FEP) face explanatory bottlenecks. We propose the Integrated Predictive Workspace Theory (IPWT), which integrates PCT/FEP as its dynamic engine, WT as its architecture, and functionally reconstructs IIT. IPWT redefines "integration" from IIT's "physical causal irreducibility" to the information-theoretic "logical irreducibility of synergistic information" (Ω_t), aligning with empirical evidence like Luppi et al.'s (2024) revised Φ value (Φ_R). We introduce Predictive Integrity (PI) and its integral ($\int$PI) as operational proxies, arguing for their convergence with high Ω_t under real-world constraints. IPWT offers a unified computational account for diverse conscious states (e.g., blindsight, schizophrenia, psychedelic states) and a "functional labeling" solution for Qualia. IPWT reconciles and unifies GWT, IIT, and FEP, providing new insights for neuroscience, clinical applications, and AGI development.

Keywords Consciousness · IPWT · Predictive Coding · Integrated Information Theory · Global Workspace Theory · Free Energy Principle · Qualia · Synergistic Information · Predictive Integrity · Neurobiology · Artificial General Intelligence

A Note on Provenance and Identity

IPWT posits that a conscious system's reality is defined by its verifiable information, not its physical carrier. We apply this principle to the theory itself. Rui Lin'' is a deliberate persona; the author's identity is irrelevant to the theory's logical integrity. Its proof-of-existence is immutably timestamped on GitHub.

A self-consistent system is its own final proof.

1 Introduction: Challenges in Consciousness Science and the Necessity of a Unified Framework

Consciousness, as the most direct yet elusive phenomenon of human experience (Block 2002), constitutes the core "hard problem" in science and philosophy (Chalmers 1995; 2007). Although neuroscience has made significant progress in identifying neural correlates of consciousness (NCCs) associated with specific conscious states over the past few decades (Seth, Baars, and Edelman 2005; Boly et al. 2013), such as localizing specific brain region activities related to visual perception, pain, or self-awareness (Koch et al. 2016), these findings are essentially correlational. We still lack a universally accepted unified theoretical framework regarding **how** consciousness truly **emerges** (Kim 1999) from the brain's complex biophysical system, how its rich phenomenological features—such as ineffable subjective qualia, the unity of experience, and its indivisibility (integration)—are formed, and what the precise functional role of consciousness is in cognitive activities.

Currently, the field of consciousness science faces a "Tower of Babel" dilemma: multiple theories coexist, but they lack deep dialogue and integration. Mainstream theories such as Integrated Information Theory (IIT), Global Workspace Theory (GWT), and Predictive Coding Theory (PCT)/Free Energy Principle (FEP) each offer profound insights into one or more aspects of consciousness from different perspectives, such as the intrinsic causal structure of information integration, the global accessibility of information broadcasting, and the minimization of prediction errors in Bayesian inference. However, these theories also face severe theoretical challenges and practical limitations. For example, IIT is criticized for its computational complexity and strong dependence on physical substrates (Toker and Sommer 2019); GWT struggles to explain the origin of subjective qualia (Seth and Bayne 2022); and PCT/FEP needs to clarify the precise link between its predictive processing mechanisms and subjective experience (Hohwy 2020).

This theoretical fragmentation not only hinders our holistic understanding of the nature of consciousness but also limits the effective translation of basic research into clinical applications. For instance, when dealing with patients suffering from schizophrenia, dissociative identity disorder, or disorders of consciousness, a unified theoretical framework would better guide our understanding of their pathological mechanisms and the development of more targeted treatment strategies. Therefore, constructing a unified framework that integrates the strengths of various theories, compensates for their shortcomings, and is more comprehensive and explanatory has become an urgent and necessary intellectual task. The Integrated Predictive Workspace Theory (IPWT) proposed in this paper aims to perform a **deep computational reconstruction and creative functional integration** of the core insights from existing theories, hoping to promote paradigm integration in consciousness science and provide a new starting point for understanding the mysteries of the human mind.

1.1 Historical Development and Latest Advances in Mainstream Theories of Consciousness

Consciousness science, as an independent and rigorous interdisciplinary field, is relatively young but has developed rapidly. After experiencing a "winter" dominated by behaviorism for most of the 20th century, scientific research on consciousness saw a resurgence at the turn of the century. This revival benefited from the rise of cognitive science, the rapid development of neuroimaging

technologies, and the introduction of tools from theoretical physics and information theory. To better understand the theoretical positioning and contributions of IPWT, it is necessary to first review this challenging and breakthrough-filled development process.

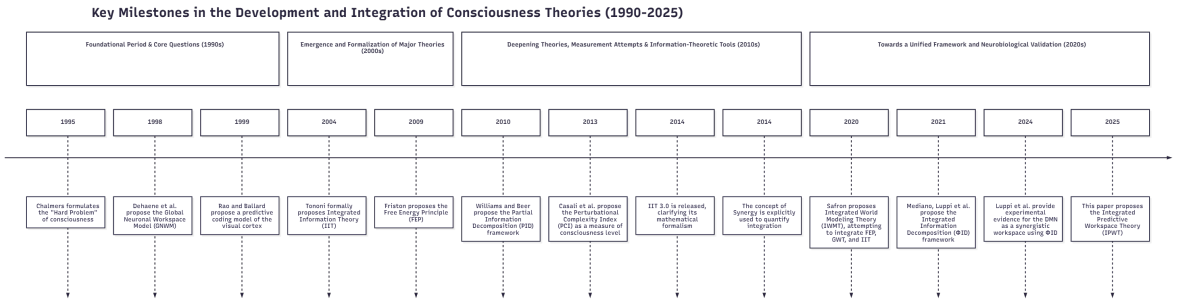

Figure 1: Key Milestones in the Development and Integration of Consciousness Theories (1990-2025)

As shown in the figure above, the 1990s marked the foundational period of consciousness science. Philosopher David Chalmers clearly distinguished between the "easy problems" of consciousness —explaining how cognitive functions are implemented—and the "hard problem"—explaining why and how subjective experience arises, setting the core agenda for the entire field (Chalmers 1995). Almost simultaneously, functional explanatory frameworks began to emerge, such as Bernard Baars' Global Workspace Theory (GWT) (Baars 1988; Baars and Franklin 2007), which was developed into its neuroscience version—the Global Neuronal Workspace Model (GNWM) —by Stanislas Dehaene et al. (Dehaene, Kerszberg, and Changeux 1998), while predictive coding (PCT) also began to appear as a theory explaining cortical processing mechanisms (Rao and Ballard 1999; Dayan et al. 1995).

Entering the 21st century, two highly influential theoretical systems—Integrated Information Theory (IIT) (Tononi 2004) and the Free Energy Principle (FEP) (Friston 2010) —emerged successively, providing more profound and formalized explanations for consciousness from the perspectives of the intrinsic causal structure of information integration and the system dynamics of Bayesian inference, respectively. In the 2010s, theoretical research further deepened. On one hand, information-theoretic tools, such as Partial Information Decomposition (PID) (Williams and Beer 2010) and Synergy (Griffith 2014), were introduced to quantify the nature of information integration more precisely; on the other hand, attempts to objectively measure conscious states also achieved breakthroughs, such as the Perturbational Complexity Index (PCI) based on transcranial magnetic stimulation (TMS) (Casali et al. 2013; Owen et al. 2006).

Entering the 2020s, with the maturation of computational neuroscience, the focus of research began to shift towards the integration and validation of existing theories. The introduction of Integrated Information Decomposition (ΦID) (Luppi et al. 2021) provided an unprecedentedly powerful tool for quantifying information integration in dynamic systems and quickly gained critical neurobiological validation in 2024 through research by Luppi et al., who found that the Default Mode Network (DMN) plays the role of a "synergistic information gateway" in conscious states (Andrea I. Luppi, Mediano, et al. 2024). These theoretical and experimental advances collectively paved the way for our proposal of the Integrated Predictive Workspace Theory (IPWT) today.

Below, we will provide a more detailed overview of the latest advances and core challenges of these mainstream theories to reveal their respective contributions and limitations, and to clarify how IPWT integrates and innovates upon them.

1.2 Latest Advances and Challenges of Integrated Information Theory (IIT)

Integrated Information Theory (IIT), first formally proposed by Giulio Tononi in 2004 (Tononi 2004), has undergone continuous iterations over nearly two decades, aiming to provide a principled, physics-based scientific explanation for the fundamental phenomenon of consciousness. IIT starts from phenomenology itself: it first extracts the undeniable core properties that any conscious experience must possess (axioms), and then, from these axioms, derives the conditions that the physical substrate supporting these experiences (postulates) must satisfy. Its core argument is that consciousness is identical to a system's ability to integrate information; a physical system is conscious if and only if its causal structure can specify a "conceptual structure" in an "integrated" manner, and the degree of this integration can be measured by a precise quantitative index—Φ (Phi) value (Oizumi, Albantakis, and Tononi 2014; Tononi et al. 2016; Tononi 2015; Kleiner and Tull 2021).

The latest version of IIT, **IIT 4.0** (Albantakis et al. 2023; Tononi et al. 2025), further refines and formalizes its theoretical framework. It starts from five phenomenological axioms —intrinsic existence, composition, information, integration, and exclusion—and derives five corresponding postulates that the physical substrate must satisfy. IIT 4.0 introduces more precise mathematical tools to evaluate the causal structure of a system, aiming to uniquely determine the "conceptual structure" specified by the system (i.e., the Qualia space) and calculate its irreducibility (Φ value). This theory not only attempts to answer "whether" a system is conscious and "how much" consciousness it has, but also "what" its conscious experience is like. In practice, IIT has led to clinical measurement methods such as the Perturbational Complexity Index (PCI), which objectively measures consciousness levels by assessing the complexity of the brain's response to external perturbations (e.g., TMS), showing great potential for application in patients with clinical disorders of consciousness (Casali, Gosseries, Rosanova, Boly, Sarasso, Casali, Casarotto, Bruno, Laureys, Tononi, and Massimini 2013; Sarasso et al. 2014).

However, despite IIT's significant theoretical progress and some empirical success, it continues to face a series of profound challenges from both the scientific and philosophical communities:

1. **Computational Infeasibility and Scalability Issues of Φ Value**: For any complex system of even moderate size (e.g., the human brain), precisely calculating its core metric Φ value is an NP-Hard problem (Toker and Sommer 2019). This means that directly applying IIT's complete mathematical framework to whole-brain neural data is computationally infeasible. Although researchers are continuously exploring various approximate computational methods, this huge computational gap largely prevents direct and complete testing of IIT's core predictions at the macroscopic level, limiting its direct applicability as an empirical science (Aguilera 2019).
2. **Strong Binding to Physical Substrate and "Substrate Independence" Controversy**: A core claim of early IIT versions was that consciousness is tightly linked to the "intrinsic causal structure" of specific physical systems, particularly its assumption of "physical causal irreducibility." This led to a controversial inference: any functionally

equivalent system (e.g., a computer program perfectly simulating the human brain) that is physically implemented differently might not possess the same conscious experience as the human brain, or even no consciousness at all (Koch 2019; Kleiner and Ludwig 2024). This strong binding to specific physical substrates stands in stark contrast to the widely held "substrate independence" or functionalist view in artificial intelligence and cognitive science. IPWT precisely attempts to resolve this core contradiction by reconstructing "physical causal irreducibility" into "logical irreducibility of synergistic information."

3. **Controversy over the Nature of Qualia and Explanatory Gap**: Although IIT claims that its "conceptual structure" is mathematically equivalent to the Qualia space of phenomenal experience (Balduzzi and Tononi 2009), this claim is far from universally accepted. Critics argue that the Φ value itself, as a scalar, primarily measures the "quantity" (intensity or degree) of consciousness, rather than its "quality" (content or feeling). Whether IIT truly explains the "what-it-is-likeness" of subjective qualia, or merely re-describes its structure, remains an unresolved philosophical question (Mørch 2019; Negro 2023; Mallatt 2021; Kelly 2022).

4. **Neglect of Dynamism and Functionality, and Challenges from Adversarial Experiments**: IIT focuses more on the static causal structure of a system at a given moment, while its explanatory power for the dynamic fluidity of consciousness and its specific functional role in guiding adaptive behavior of organisms is relatively weak. In recent years, IIT and GWT have directly confronted each other in a large-scale "adversarial collaboration" project, aiming to test the conflicting predictions of the two theories through a series of carefully designed experiments (Melloni et al. 2023). Some preliminary published research results, for example, on the sustained representational role of posterior cortex in conscious perception, seem to challenge some core predictions of IIT, indicating that real neural dynamics are more complex than the theory presupposes (Andrea I. Luppi, Mediano, et al. 2024).

5. **"Pseudoscience" Controversy and Debate on Scientific Status**: In 2025, over a hundred scientists jointly published an open letter accusing IIT of being "pseudoscience" due to some of its inferences (e.g., panpsychist tendencies) and the unfalsifiability of its core claims. This fierce debate quickly sparked a major discussion in academia about "what constitutes a scientific theory" and "how to test theories of consciousness" (Gomez-Marin and Seth 2025). Supporters of IIT responded that the theory makes numerous testable predictions, and its counter-intuitive conclusions should not be a reason for rejection, but rather a reflection of its theoretical depth (Tononi, Albantakis, Barbosa, Boly, Cirelli, Comolatti, Ellia, Findlay, Casali, Grasso, Haun, Hendren, Hoel, Koch, Maier, Marshall, Massimini, Mayner, Oizumi, Szczotka, Tsuchiya, and Zaeemzadeh 2025; Klincewicz et al. 2025; Kastrup 2023; Guerrero et al. 2025). This debate highlights the fundamental difficulties that consciousness science still faces in theoretical construction and experimental validation paradigms.

1.3 Latest Advances and Challenges of Global Workspace Theory (GWT)

Unlike IIT, which starts from phenomenological axioms and intrinsic causal structure, Global Workspace Theory (GWT) offers a more functionalist and cognitively oriented model of consciousness. GWT was first proposed by Bernard Baars in the late 1980s (Baars 1988), and

its core idea is highly inspiring: he likened the function of consciousness to a theater stage. In this analogy, the cognitive system consists of a vast number of parallel, unconscious specialized processing modules working silently "offstage." At any given moment, only information selected by the "spotlight" of attention enters a limited-capacity "global workspace" (the stage) and is then **globally broadcast** to all "audiences" (other specialized modules) throughout the cognitive system (Baars 2002). Once information is broadcast, it becomes "conscious" information, capable of flexibly guiding behavior, enabling verbal reports, and forming episodic memories.

GWT clearly elucidates the functional role of consciousness in information processing, cognitive regulation, and behavioral control, and successfully explains several key features of conscious experience, such as limited capacity (we can only be aware of a few things at a time), sequentiality (conscious content appears in temporal order), and information integration and sharing. Its neuroscience version—the **Global Neuronal Workspace Model (GNWM)**—proposed by Stanislas Dehaene and Jean-Pierre Changeux, suggests that consciousness arises from the "ignition" of a widely distributed cortical network system composed of long-range connected pyramidal neurons in the brain (Dehaene, Kerszberg, and Changeux 1998; Mashour et al. 2020). When the strength and duration of an information representation are sufficient to trigger a nonlinear, self-amplifying activation of this network, the information becomes globally available, thereby generating subjective conscious experience.

In recent years, GWT/GNWM theory has made significant progress in theoretical deepening, neural mechanism elucidation, and application expansion, while the challenges it faces have also prompted continuous refinement of the theory:

1. **Theoretical Deepening and Dynamization**: GWT has evolved from a relatively static architectural model to a "Global Workspace Dynamics (GWD)" (Baars, Franklin, and Ramsoy 2013; Baars and Geld 2019) that emphasizes dynamic processes. This view highlights the dynamic and oscillatory properties of the cortico-thalamic (C-T) system, treating it as a "unified oscillatory machine" that transcends fixed anatomical divisions and moves towards a more integrated view of overall cortical function. This dynamic perspective suggests that consciousness arises from "binding and propagation" processes within cortical networks, rather than merely the activity of specific brain regions.
2. **Applications in Artificial Intelligence (AI) and Artificial General Intelligence (AGI)**: GWT's architectural ideas provide a blueprint for building more advanced AI systems. Recent research has explored the possibility of explicitly implementing GWT in deep learning and AGI (Rullen and Kanai 2020; Devillers, Maytié, and VanRullen 2025; Shanahan 2006; Huang, Chella, and Cangelosi 2024). For example, by mimicking GWT's information bottleneck and broadcasting mechanisms, researchers have proposed concepts such as "Global Latent Workspace (GLW)," aiming to enhance the generality and multimodal integration capabilities of AI models by allowing multiple specialized AI models to share a common representational space (Abdelwahab, M., and & Aarabi 2023; Dossa et al. 2024).
3. **Explanation of Subjective Qualia**: While GWT primarily focuses on the "function" rather than the "feeling" of consciousness, recent research has begun to attempt to bridge this gap. Some evidence suggests that, contrary to traditional views, the prefrontal cortex (PfC)—a core region of GNWM—may be directly involved in the formation of sensory conscious experience, including its subjective qualia (Fox et al. 2020). This challenges the view that

PfC function is strictly limited to high-level cognitive control and suggests a more direct link between global broadcasting processes and the generation of subjective experience.

4. **Clarification of Specific Mechanisms and Boundary Issues**:
   - **Information Selection and Broadcasting**: The core concept of "ignition" is further elucidated as a bidirectional information broadcasting process regulated by attentional mechanisms within the cortico-thalamic system.
   - **Neural Implementation Mechanisms**: Mathematical models such as "Cortical Neuropercolation (CNP)" have been proposed to describe how cortical networks transition from fragmented local activity states to globally coherent activity states through phase transitions, providing a dynamic description of how information achieves global accessibility (Kozma and Freeman 2017).
   - **Temporal Dynamics**: Research has revealed that conscious access has discrete temporal dynamic characteristics, with one "ignition" and broadcasting process taking approximately 100-300 milliseconds, which contrasts sharply with the speed of unconscious, automated processing (Dehaene and Changeux 2011).
5. **Enhanced Explanatory Power for Complex Conscious States**: The GWT framework has been successfully applied to explain various complex conscious states. For example, **metacognition** is considered the workspace's monitoring of its own state (Baars, Geld, and Kozma 2021; Lau and Rosenthal 2011); **dreaming** is explained as workspace activity driven by endogenous information in the absence of external sensory input, while **lucid dreaming** is related to the restoration of metacognitive function during dreaming; **meditation** has been found to functionally reorganize workspace activity patterns, particularly by altering the involvement of the Default Mode Network (DMN), thereby enhancing cognitive flexibility (Brewer et al. 2011); **hypnosis** is related to selective changes in workspace function, leading to the dissociation of perception, memory, and action control.

Despite these advances, GWT continues to face criticism and challenges. For example, the debate continues regarding the precise roles of the prefrontal cortex and posterior cortex in the generation of consciousness, but the trend favors a more dynamic and integrated view. Furthermore, the methodology of some classic experimental paradigms used to support unconscious processing capabilities (e.g., unconscious priming) has also been questioned (Meyen et al. 2022). Most importantly, GWT still needs to provide more precise and operationalizable definitions for its core mechanisms—how information is "selected" to enter the workspace, and what exactly "broadcasting" entails as a neural process.

1.4 Unified Explanatory Power and Latest Advances of Predictive Coding (PCT) and Free Energy Principle (FEP)

Predictive Coding Theory (PCT) and the Free Energy Principle (FEP) together constitute one of the most influential and unifying theoretical frameworks in contemporary cognitive neuroscience. PCT was initially proposed by Rao and Ballard in 1999 to explain information processing mechanisms in the visual cortex (Rao and Ballard 1999). Its core idea is that the brain does not passively receive and process sensory information, but is an active **prediction machine** (Clark 2013; Friston 2018). Higher-level brain regions continuously generate predictions about lower-level sensory inputs (top-down predictive signals), while lower-level regions are responsible for comparing these predictions with actual sensory inputs and transmitting the mismatch between the two—i.e., **prediction error**—upwards. This bottom-

up error signal is then used to revise higher-level predictions, forming a continuous perceptual loop aimed at minimizing prediction error.

Subsequently, Karl Friston generalized and extended the core ideas of PCT, developing the more universal **Free Energy Principle (FEP)** (Friston 2010; Friston 2005; Parr, Pezzulo, and Friston 2022; Friston, Kilner, and Harrison 2006). FEP states that any self-organizing system capable of resisting entropy increase (from a single cell to the entire brain) must minimize its **variational free energy** through its actions and states. Variational free energy is an information-theoretic measure that quantifies the degree of mismatch between the predictions of the system's internal generative model and the true state of the external world, essentially an upper bound on "surprise." Therefore, FEP unifies the brain's function into a single goal: minimizing free energy. The system can achieve this goal in two ways: **changing its internal model to better fit sensory input (perceptual inference and learning)**, or **changing sensory input through action to make it more consistent with predictions (active inference and action)** (Friston et al. 2010).

The PCT/FEP framework, with its immense unifying explanatory power, has successfully placed perception, learning, attention, motor control, and even various symptoms of mental illness under the same mathematical and computational framework (Adams, Shipp, and Friston 2013; Sterzer et al. 2019). In recent years, FEP has been further established as "one of the most encompassing ideas since Darwin's theory of natural selection," aiming to provide a unified principle for life, mind, and intelligence (Georgiev 2025; Gong 2024).

Although the PCT/FEP framework has made significant theoretical progress, its direct theoretical bridge to subjective conscious experience is still under construction and faces the following challenges and latest advances:

1. **Emergence of Conscious Content and Qualia Explanation**: Traditionally, PCT/FEP has focused more on explaining the "how" of cognitive processes rather than the "why" of subjective experience. However, recent theoretical developments have begun to directly address this challenge. Anil Seth proposes that emotions and subjective feeling states (Qualia) are precisely generated by predictive models that specialize in predicting and regulating interoceptive signals from within the body (Seth 2013; Barrett and Simmons 2015). Lisa Feldman Barrett's "theory of constructed emotion" further elaborates on how emotional experiences are actively "constructed" through the interaction of interoceptive predictions and conceptual categories (Barrett 2017). These advances provide a functional, prediction-based explanation for Qualia, suggesting that subjective experience is the system's optimal prediction and control of its own physiological state and interaction with the environment, thereby transforming the "hard problem" into a scientifically tractable problem of interoceptive inference (Solms 2019).

2. **Unity and Boundaries of Consciousness**: The PCT/FEP framework, through its hierarchical generative models, provides a natural explanation for the integration of multimodal information. Information from different sensory modalities can be integrated at higher levels of the model to produce a unified, coherent world model. Furthermore, through the formalized concept of a "Markov blanket," FEP provides a statistical definition for the distinction between "self" and "non-self" (Friston 2013; Farnsworth 2025). A Markov blanket is a statistical boundary that separates a system's internal states from its environment, allowing the system to infer and interact with the external world only through its "blanket" (i.e., sensory and motor states). This provides a principled, system-

environment boundary-based perspective for understanding the formation and maintenance of self-awareness.

3. **New Experimental Evidence, Computational Models, and Clinical Applications**: PCT/FEP's predictions have been validated in multiple experimental paradigms. For example, phenomena such as repetition suppression and mismatch negativity (MMN) in EEG signals have been successfully explained as manifestations of prediction error signals (Garrido et al. 2009). In the field of artificial intelligence, FEP and Active Inference ideas are widely applied to build more autonomous and adaptive reinforcement learning agents and world models for Artificial General Intelligence (AGI) (Maier 2025; PubMed 2024; Constant et al. 2022). Additionally, new software frameworks (e.g., `RxInfer`) are being developed to facilitate researchers in building and testing FEP-based computational models.

4. **New Criticisms or Challenges**: Although FEP has immense explanatory power, its vast universality also brings some problems. Critics point out that because FEP is a "principle" rather than a specific "theory," it sometimes struggles to generate sufficiently precise, falsifiable predictions (Hohwy 2020; Radomski and Dołęga 2024). Furthermore, the specific neural computational methods for minimizing prediction error, such as how error signals are precisely weighted and transmitted, remain controversial. Finally, the computational tractability problem still exists: while PCT/FEP is conceptually elegant, the Bayesian inference involved at each level can be computationally very complex or even intractable, posing a challenge to its biological plausibility in the real brain.

1.5 The Proposal of IPWT: Deep Integration and Neurobiologically Driven Reconstruction

By reviewing the three major theories—IIT, GWT, and PCT/FEP—we can clearly see their respective brilliant achievements and unresolved problems. IIT offers profound phenomenological insights and attempts at mathematical formalization for the "integrated" nature of consciousness, but it is constrained by computational bottlenecks and rigid dependence on physical substrates. GWT provides an intuitive architectural model for the "broadcast" function of consciousness and its role in cognitive regulation, but it is relatively weak in explaining the origin of subjective experience. PCT/FEP, on the other hand, offers powerful unifying computational principles for the "generation" of conscious content and the dynamic processes of the brain, but its direct link to subjective consciousness still requires clearer elucidation.

In recent years, there have been some attempts in academia to integrate these theories, such as Safron's Integrated World Modeling Theory (IWMT) (Safron 2020; 2022), which attempts to unify IIT and GWT within the FEP framework. These attempts are far-sighted, correctly recognizing that the future of consciousness science lies in theoretical convergence rather than continued fragmentation. However, these early integration models often failed to provide an intrinsically consistent, computationally feasible, and broadly explanatory unified framework, particularly failing to fundamentally address the core challenge faced by IIT: how to "liberate" its profound insights into "integration" from its controversial physical and computational assumptions.

It is against this academic backdrop that we propose the **Integrated Predictive Workspace Theory (IPWT)**. The core objective of IPWT is not merely a simple "patchwork" or "piecing together" of existing theories, but rather an attempt to construct a new, intrinsically consistent,

and powerfully externally explanatory unified framework of consciousness through a **deep computational reconstruction and creative functional integration** of the core insights from PCT/FEP, WT, and IIT.

IPWT's integration is **selective, asymmetric, and functionally driven**:

- We adopt **PCT/FEP as the dynamic foundation of the entire framework**, positing that conscious processes are inherently prediction-driven.
- We adopt and extend **WT as the architectural platform for information integration and broadcasting**, but generalize it from a single "global" space to more flexible, dynamically generatable "Workspace Instances" (WSIs).
- We perform a **fundamental functional reconstruction of IIT's core contribution**: we retain its phenomenological insight that "integration" is a core feature of consciousness, but decisively abandon its reliance on "physical causal irreducibility," instead redefining integration using the more universal, flexible, and computationally friendly concept of "logical irreducibility of synergistic information" from information theory.

In this way, IPWT aims to build a unified model that retains the strengths of each theory while overcoming their core drawbacks. In the following chapters, we will elaborate on IPWT's theoretical framework, its core computational reconstruction, operationalizable measurement methods, and how it provides a new, neurobiologically driven perspective for explaining various conscious phenomena, from normal to abnormal.

2 The IPWT Framework: Mechanistic Emergence of Consciousness

The Integrated Predictive Workspace Theory (IPWT) aims to construct a unified framework for how consciousness mechanistically emerges from neural activity by integrating the core mechanisms of Predictive Coding Theory (PCT), the Free Energy Principle (FEP), and Workspace Theory (WT), and by functionally reconstructing the phenomenological axioms of Integrated Information Theory (IIT). This chapter will detail the core components of IPWT, their interactions, and how they collectively form a coherent and explanatory model of consciousness.

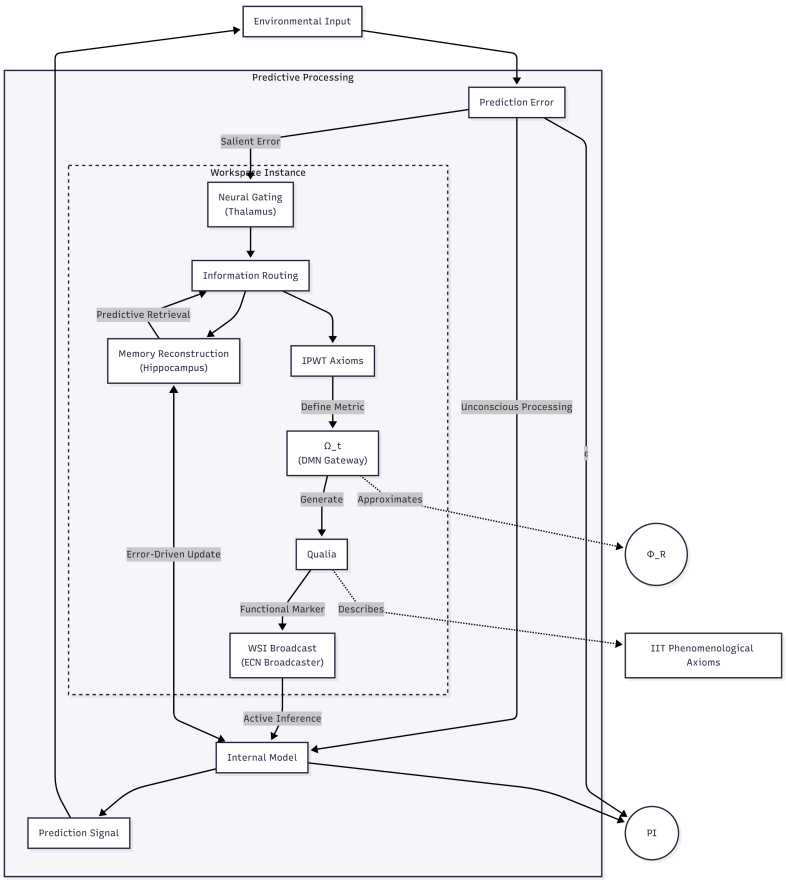

Figure 2: The Core Logic of the IPWT Framework

The figure above visually illustrates the core logic of the IPWT framework. The entire cognitive system is viewed as a **PCT/FEP-based predictive processing core**, which continuously interacts predictively with the environment through an internal generative model. When prediction errors are sufficiently "salient," this information is fed into one or more dynamically formed **Workspace Instances (WSIs)**. Within the WSI, information undergoes **integration**, forming a synergistic, logically irreducible whole (its degree of integration is measured by Ω). This integrated information is then **selectively broadcast** to the entire system, used to update the internal generative model (learning) and guide the organism's next actions (active inference). This process is cyclical and dynamic, with conscious content continuously emerging and updating within this prediction, integration, and broadcasting loop.

2.1 Basic Assumptions and Core Principles

The IPWT theoretical framework is built upon the following four interconnected core assumptions, which collectively form the basis for our understanding of how consciousness emerges from complex neural computations.

1. **Assumption One: Consciousness is an emergent phenomenon of information integration and synergistic processing.** We adopt a clear functionalist stance, positing that consciousness is not some mysterious, non-physical attribute, but rather a complex phenomenon that emerges when information is efficiently integrated and synergistically processed within a specific functional architecture. The core here is **synergy**—when multiple independent information units combine, they can produce a novel causal effect and semantic meaning that cannot be reduced to the independent contributions of their parts. This view

implies that, in principle, any system capable of achieving the same functional architecture and information dynamics, regardless of its physical substrate (biological neurons or silicon chips), could potentially generate consciousness. This provides a theoretical basis for the substrate independence of consciousness.

2. **Assumption Two: Consciousness is a prediction-driven, dynamic process aimed at minimizing free energy.** Inspired by the PCT/FEP framework, we view consciousness as a continuous, prediction-centric **dynamic process**, rather than a series of discrete, static **snapshots**. The cognitive system actively and continuously constructs a hierarchical generative model of the world (including both the external environment and its own body). The core function of this model is **prediction**—from high-level abstract concepts to low-level concrete sensory details, the system constantly generates top-down predictive signals, attempting to anticipate the next moment's sensory input.

   When actual sensory inputs (whether from eyes, ears, or within the body) arrive, they are compared with corresponding predictive signals. The mismatch between the two, i.e., **prediction error**, constitutes the key information driving the entire system's learning and updating. These error signals propagate bottom-up through the processing hierarchy, with their core role being to "inform" higher-level models: "Your prediction is wrong and needs correction."

   The entire process adheres to the **Free Energy Principle** (Friston 2010). The system instinctively and continuously adjusts itself to minimize the long-term average prediction error, weighted by precision. This minimization process is achieved through two complementary mechanisms:

   1. **Perceptual Inference and Learning**: When faced with prediction errors, the system can reduce future errors by **optimizing its internal generative model**. This corresponds to what we commonly call "perception" and "learning." For example, upon seeing an unexpected object, the system updates its model of the current scene to better explain the object's presence. Long-term, continuous model updates constitute knowledge acquisition and memory formation.
   2. **Active Inference and Action**: In addition to changing the model, the system can also **change sensory input through action** to make it more consistent with existing predictions. For example, if my prediction of an object's distance is inaccurate, I can reach out and touch it, eliminating prediction error by actively acquiring new sensory evidence (touch and proprioception). Thus, action itself is redefined as an inference process, also aimed at minimizing free energy (Friston, Daunizeau, Kilner, and Kiebel 2010).

   Within this framework, **conscious content is no longer seen as a direct "snapshot" of the external world, but rather the internal model's active, constructive best prediction and explanation of the internal and external world.** Information that ultimately enters conscious experience typically has the following characteristics: it is either highly consistent with high-confidence (high-precision) predictions, or it represents significant prediction errors that cannot be easily explained by the current model, and this information is highly relevant for guiding the organism's next actions.

   Furthermore, IPWT closely links the **memory system** with the internal generative model. We argue that the internal generative model is essentially a **dynamic, predictive memory system** (Rolls 2024; Mongillo, Barak, and Tsodyks 2008; Butola et al. 2023; Budson,

Richman, and Kensinger 2022; Damasio 1989). The process of memory encoding is the process of optimizing model parameters through learning; while memory retrieval is an active, prediction-driven process triggered by current contextual cues, where the system "re-enacts" or "activates" past states to predict current and future events. This view transforms memory from a static "storage warehouse" into a dynamic cognitive tool serving prediction and action.

3. **Assumption Three: Consciousness is a workspace-based information processing hub.** Building upon and extending GWT, we hypothesize the existence of one or more dynamic, limited-capacity **Workspace Instances (WSIs)** within the cognitive system. A WSI is not a fixed anatomical structure but a functional concept, referring to a group of neurons that are tightly coupled and synergistically active within a specific time window, collectively forming a critical node for information selection, integration, amplification, and broadcasting. The limited capacity of a WSI explains the "bottleneck" characteristic of conscious experience (we can typically only clearly be aware of a few things); its integration function is the basis for generating unified, coherent experiences; and its broadcasting property explains how conscious content can be utilized by other cognitive modules within the system (e.g., memory, language, motor control modules) to achieve flexible, globally coordinated cognitive functions (Baars, Geld, and Kozma 2021).

   We consider the global information broadcasting phenomenon described by GWT as a **special and highly integrated configuration state of WT**. When a WSI's integration scope is extremely broad and its integration degree is extremely high, capable of covering and influencing almost the entire cognitive system, it plays the role of the "global workspace" in traditional GWT. However, we believe that not all conscious experiences must reach this level of global integration. For example, some local, faint conscious sensations might only involve a smaller, less integrated WSI. This perspective allows IPWT to better accommodate various conscious states of different intensities and contents, from vague background feelings to clear focal awareness, thereby explaining the rich diversity of conscious experience.

   Crucially, WSIs do not exist independently of PCT/FEP's hierarchical dynamics but emerge as top-level integration hubs within this dynamic process. Their boundaries are not defined by fixed anatomical structures but are dynamically shaped "software-defined boundaries" by the computational demands of the current cognitive task. Within the system's predictive hierarchy, when lower-level modules cannot effectively suppress prediction errors through local circuits, these salient, unresolved error signals propagate upwards through the hierarchy. The function of a WSI is to act as a higher-order integration platform, selectively receiving these error signals and synergistically integrating them with contextual information retrieved from memory. This process aims to generate a more comprehensive and abstract generative model, thereby enabling the generation of new, higher-order predictions. These new predictions are then broadcast to the entire cognitive system, regulating and constraining lower-level processing in a top-down manner, with the ultimate goal of minimizing the system's long-term average prediction error, i.e., variational free energy. Therefore, the formation, operation, and dissipation of WSIs are themselves a mechanistic emergence of the system in its pursuit of maximizing predictive integrity (PI).

   Recent neurobiological discoveries provide decisive empirical support for IPWT's conception of WSIs, particularly the idea of functional heterogeneity within them. Andrea Luppi and colleagues' (2024) groundbreaking research, utilizing Integrated Information Decomposition

(ΦID), an advanced information-theoretic tool, conducted an in-depth analysis of resting-state functional MRI data, thereby mapping out a new picture of brain information processing architecture (Andrea I. Luppi, Mediano, et al. 2024).

The core finding of this research is the existence of a "synergistic global workspace" in the brain characterized by synergistic information processing. This space exhibits clear functional specialization:

- **Gateways**: These regions largely overlap with nodes of the **Default Mode Network (DMN)** (Raichle 2015). Their function is to "collect" and "integrate" **synergistic information** from various specialized modules of the brain. The DMN is anatomically and functionally ideally positioned to integrate multimodal information from different cognitive systems, consistent with its role as the primary entry point for information into the synergistic workspace (Raichle 2015; Andrea I. Luppi, Singleton, et al. 2024; Andrea I. Luppi, Girn, et al. 2024).
- **Broadcasters**: These regions are primarily located in the **Executive Control Network (ECN)**, especially the lateral prefrontal cortex. Their function is to widely "broadcast" the integrated information within the workspace, in the form of **redundant information**, to the entire brain to guide subsequent processing. This finding is highly consistent with GWT's classic view that the prefrontal cortex plays a central role in the global broadcasting of conscious content (Dehaene, Kerszberg, and Changeux 1998; Mashour, Roelfsema, Changeux, and Dehaene 2020).

Based on these findings, we introduce the concept of **Dominant Neural Workspace Instance (DNWSI)**, which explicitly bridges and distinguishes our more flexible, dynamic WSI concept in the IPWT framework from the classic ideas of GWT/GNWM. In the IPWT framework, DNWSI is precisely the functional entity that emerges from the dynamic synergistic interaction of DMN gateways and ECN broadcasters. It constitutes a more refined and dynamic neurocomputational equivalent of the classic Global Neuronal Workspace Model (GNWM) concept. DNWSI is not only responsible for global information broadcasting but, more importantly, it first achieves deep integration of synergistic information through the DMN gateways, which forms the core of subjective conscious experience. This definition not only tightly links our theory with the latest empirical evidence but also clearly elucidates the distinct yet complementary functional roles of the DMN and ECN in the generation of consciousness, deepening the theoretical implications of IPWT.

4. **Assumption Four: The integrality of consciousness arises from the logical irreducibility of synergistic information.** This is IPWT's most core theoretical innovation and our key to functionally reconstructing IIT. We fully agree with IIT's phenomenological insight that "integration" is the most fundamental and core characteristic of consciousness. However, we reinterpret its origin not from IIT's emphasis on the "physical causal irreducibility" of the physical substrate, but from the **logical irreducibility of synergistic information** formed between information units processed within the Workspace Instance (WSI). This means that when multiple independent information units (e.g., information about an object's shape, color, motion, and sound) are integrated within a WSI, the semantic meaning and causal influence of this integrated whole representation cannot be fully explained or reconstructed by simply decomposing it back into its original, isolated information units of shape, color, sound, etc. This whole produces an emergent effect where "the whole is greater than the sum of its parts."

Fortunately, this concept of "logical irreducibility" finds its precise, quantifiable mathematical counterpart in modern information theory. The **Partial Information Decomposition (PID)** framework, first proposed by Williams and Beer in 2010 (Williams and Beer 2010), can precisely decompose the total information (i.e., mutual information $I(X_1, X_2; Y)$) provided by multiple sources (e.g., X_1, X_2) about a target (Y) into four non-negative atomic parts: **unique information** provided only by X_1, unique information provided only by X_2, **redundant information** provided jointly by both, and **synergistic information** that can only be obtained when both are considered as a whole.

Among these, **synergistic information (Synergy)** precisely quantifies the emergent information where "the whole is greater than the sum of its parts." It represents new information arising from the interaction between sources that is not present in any single source. Therefore, synergistic information becomes the theoretical cornerstone used by IPWT to define and measure information integration (Griffith 2014; Bertschinger et al. 2014; Tian and Shamai 2025; Sherrill et al. 2021).

Building on this, the theoretical frontier has further developed the **Integrated Information Decomposition (Integrated Information Decomposition, ΦID)** framework (Luppi, Mediano, Rosas, Harrison, Carhart-Harris, Bor, and Stamatakis 2021). ΦID extends PID from a static "multi-source to single-target" analysis to analyzing the complete information flow from "multiple sources to multiple targets" in dynamic systems, enabling a more refined characterization of information generation, transfer, and modification in time-series data. The introduction of ΦID provides us with a solid, principled theoretical foundation to formalize the concept of "integration."

Crucially, Luppi et al.'s (2024) research precisely utilized the ΦID framework to develop a computationally tractable and empirically effective measure of integrated information—the **revised Φ value (ΦR)**. They demonstrated that, unlike earlier versions of Φ, ΦR resolves the paradox of potentially negative values and reliably tracks changes in consciousness levels induced by anesthesia or brain injury (Andrea I. Luppi, Mediano, et al. 2024). This work provides strong empirical evidence for our core argument—that information-theoretic functional measures can replace IIT's reliance on physical causation. It shows that the "logical irreducibility" we advocate is not only theoretically self-consistent but also practically measurable and closely related to conscious states.

By shifting the core of integrality from "physical irreducibility" to "logical irreducibility" (and linking it to synergistic information), IPWT successfully integrates IIT's phenomenological axioms into a more dynamic, computationally feasible, and substrate-independent theoretical framework. Now, we can functionally reinterpret IIT's five axioms:

1. **Existence**: In IPWT, an information state "exists" in consciousness if and only if it is activated within a WSI and exerts a sustained, measurable **functional influence** on the system's future states and behaviors (Presti 2021).
2. **Information**: Every information state that "exists" in a WSI carries unique, distinguishable **content or semantics**. It reduces uncertainty for the entire system by specifying a particular possibility among many, thereby guiding its inference and action (Celotto et al. 2023).
3. **Integration**: Multiple independent information units converge and associate within a WSI, forming a **logically irreducible, functionally unified, synergistic cognitive state**. The meaning, predictive power, and causal effects of this integrated whole

transcend the simple sum of its parts. Its degree of integration can, in theory, be quantified by synergistic information (Hancock et al. 2025).

4. **Exclusion**: Due to the limited capacity of the WSI and its internal competitive dynamics (winner-take-all), at any given moment, only one or a few of the most salient, highly integrated cognitive states that best minimize global prediction error can dominate the WSI, becoming the core content of current conscious experience. Other competing representations are excluded from consciousness (Dehaene et al. 1998).
5. **Causation**: Information states that are integrated within the WSI and achieve dominance possess significant **causal power**. Through broadcasting mechanisms, they can influence the activity of other cognitive modules within the system (e.g., updating memory, adjusting attention) and ultimately guide the organism's overt behavior and decision-making. This causal power is functional, not metaphysical (Chalmers 2011).

3 Computational Verifiability and Neurobiological Measures

A mature scientific theory must not only provide unified and profound explanations at the conceptual level but also translate its core claims into operationalizable measurement methods and testable empirical predictions. IPWT, as a theoretical framework aimed at promoting paradigm integration in consciousness science, has placed **computational verifiability** and **empirical testability** at its core from its inception. This chapter aims to elaborate on how IPWT transforms the core concept of "information integration" from an abstract philosophical idea into concrete metrics that can be computed and validated in real neurobiological data.

We will construct this bridge from theory to practice in three steps, clearly demonstrating how IPWT moves from abstract theory to empirical validation:

1. **Theoretical Gold Standard (Ω_t)**: We first define the theoretical "gold standard" measure —**instantaneous information integration (Ω_t)**. This measure is based on the concept of synergistic information and aims to precisely characterize the "logical irreducibility of information integration" from a fundamental information-theoretic perspective.
2. **Empirically Computable Proxy (Φ_R)**: Next, we introduce the Integrated Information Decomposition (ΦID) framework and elaborate on how Luppi et al.'s (2024) proposed **revised Φ value (Φ_R)** serves as a computable and empirically supported proxy for Ω_t in real neural data, thereby linking our theory to experimental measurements.
3. **Functionally Computable Proxy (PI/∫PI)**: Finally, we introduce the computationally more efficient **Predictive Integrity (PI)** and its temporal integral (∫PI) as functional proxies. We will argue that, under real-world constraints, the pursuit of high predictive performance (high PI) inevitably drives the system towards high information integration (high Ω_t), thereby establishing the effectiveness and practicality of PI/∫PI as measures of conscious states.

3.1 Instantaneous Information Integration (Ω_t): A Theoretical Definition Based on Synergistic Information

We propose that the core mechanism underlying consciousness is information integration, specifically referring to the formation of a logically irreducible, functionally unified, synergistic cognitive state from multiple independent information units within a Workspace Instance

(WSI). To transform this core idea from a philosophical concept into a scientifically operational measure, we must provide a precise, quantifiable mathematical definition. To this end, we draw upon the Partial Information Decomposition (PID) framework, particularly its core concept —**Synergistic Information (CI)**—to define a theoretical "gold standard" measure we call **instantaneous information integration (Ω_t)**.

As previously discussed, the PID framework aims to decompose the total information (i.e., total mutual information $I(X_1, ..., X_n; Y)$) provided by multiple sources $X_1, ..., X_n$ about a target Y into atomic parts such as redundant, unique, and synergistic information (Williams and Beer 2010). Among these, synergistic information (CI) refers to the "emergent" information that can only be obtained when all sources are considered as a whole; it cannot be obtained from any subset of the sources. Therefore, CI precisely captures the essence of the "logical irreducibility of information integration" that we emphasize—that part of the information where "the whole is greater than the sum of its parts" (Griffith 2014; Bertschinger, Rauh, Olbrich, Jost, and Ay 2014; Proca et al. 2024; Varley 2023).

Based on this, we theoretically define **instantaneous information integration (Ω_t)** as: the proportion of **synergistic information (CI)** generated by a set of information units $X = \{X_1, ..., X_n\}$ used to predict a target variable Y within a specific WSI, relative to the **total predictive information (i.e., total mutual information $I(X; Y)$)** it provides for predicting Y.

$$\Omega_t(X \to Y) = \frac{\text{CI } (X_1, ..., X_n; Y)}{\text{I}(X_1, ..., X_n; Y)} \tag{1}$$

The intuitive meaning of this formula is: among all information utilized by the WSI to achieve a certain function (i.e., predicting Y), what proportion is "truly integrated" and indivisible? A high Ω_t value (maximum 1) means that the information in the WSI is primarily integrated and utilized in a highly synergistic, irreducible manner, which corresponds to our intuitive understanding of a highly unified, coherent conscious state. Conversely, a low Ω_t value (minimum 0) means that the information in the WSI primarily exists in redundant or independent ways, which might correspond to a fragmented, poorly integrated conscious state, or even unconscious information processing.

Ω_t is an **idealized standard**. It provides an unambiguous, information-theoretic definition of "information integration" based on first principles. However, due to the extreme mathematical and computational difficulty of directly calculating synergistic information for high-dimensional systems (i.e., a large number of sources X_n), this makes Ω_t currently difficult to apply directly and precisely to large-scale neural data in practice.

Nevertheless, the theoretical value of Ω_t is immense. It provides a clear target against which all other proxy measures of information integration should be evaluated for their theoretical approximation to Ω_t. In this sense, we believe that IIT's proposed integrated information Φ value can be regarded as an attempt at a **physical instantiation** of IPWT's information integration measure Ω_t in specific physical systems (such as the biological brain). We hypothesize that when a system is *physically closed and its causal structure is fully known*, its IIT Φ value, calculated based on "physical causal irreducibility," is conceptually highly related to our defined Ω_t based on information flow synergy. IIT Φ measures the integration of the physical substrate's intrinsic causal capacity, while Ω_t measures the functional integration of information processing. In the specific implementation of the biological brain, both are

likely describing different aspects of the same phenomenon: an efficient information-integrating biological network must also possess a highly integrated physical causal structure.

3.2 From Ω to ΦR: Integrated Information Decomposition (ΦID) and Empirical Proxies

The theoretical gold standard Ω_t provides us with a fundamental definition of information integration, but its computational complexity limits its direct application to real neural data. To advance IPWT from theory to practice, we must find an **empirical proxy** that is both faithful to the core idea of Ω_t and computationally feasible. Fortunately, recent advances in information theory, particularly the introduction of the **Integrated Information Decomposition (ΦID)** framework (Luppi, Mediano, Rosas, Harrison, Carhart-Harris, Bor, and Stamatakis 2021), and the groundbreaking empirical research by Luppi et al. (2024) based on this framework, provide us with such a crucial bridge (Andrea I. Luppi, Mediano, et al. 2024).

The ΦID framework is a dynamic extension of Partial Information Decomposition (PID) (Williams and Beer 2010), aiming to precisely decompose the overall information influence that multiple sources $(X_1, ..., X_n)$ in past states have on their future states $(Y_1, ..., Y_m)$ in a dynamic system into three atomic parts: Redundancy, Uniqueness, and Synergy. Among these, synergistic information precisely captures the emergent effect of "the whole being greater than the sum of its parts," which is highly consistent with our defined Ω_t conceptually.

Luppi et al.'s (2024) research precisely utilized the ΦID framework to develop a theoretically sounder and empirically more effective measure of integrated information—the **revised Φ value (Φ_R)**. They first pointed out a key flaw in earlier versions of Φ: in some cases, the value could be negative, which is intuitively difficult to explain. Through ΦID decomposition, they demonstrated the source of this paradox—the original Φ calculation subtracted redundant information within the system. Therefore, they proposed a revised solution, $\Phi_R = \Phi + \text{Red}(X, Y)$, which ensures the non-negativity of the measure by adding back the redundant information.

Crucially, their research showed that this theoretically more reasonable Φ_R metric empirically tracks changes in conscious states very effectively:

1. **Clinical Validity**: In patients under anesthesia and with disorders of consciousness (DOC), the Φ_R value of the brain's synergistic workspace (especially the Default Mode Network DMN nodes acting as "gateways") significantly decreased.
2. **State Reversibility**: After recovery from anesthesia, the Φ_R value also rebounded accordingly.

This work provides strong empirical evidence for our core argument—that information-theoretic functional measures can replace IIT's reliance on physical causation. It demonstrates that the "logical irreducibility" we advocate is not only theoretically self-consistent but also practically measurable and closely related to conscious states (Andrea I. Luppi, Mediano, et al. 2024; Andrea I. Luppi, Rosas, et al. 2024). Therefore, within the IPWT framework, we consider Φ_R to be the most promising empirical proxy currently available for approximating Ω_t in real neural data. It successfully links the abstract concept in our theory (Ω_t) with a quantity that can be concretely computed and validated in neuroimaging data such as fMRI, laying a solid empirical foundation for IPWT as a true scientific theory.

3.3 Predictive Integrity (PI) and Functional Proxies

After establishing the bridge from the theoretical gold standard (Ω_t) to the empirical proxy (Φ_R), we also need a computationally more efficient and readily applicable **functional proxy metric** for various neural data. To this end, IPWT introduces **Predictive Integrity (PI)** and its temporal integral (∫PI). The core idea behind this initiative is that a system capable of efficient synergistic information integration (high Ω_t) will inevitably exhibit stronger predictive capabilities and higher state stability. Therefore, by measuring the system's performance in predictive efficacy, we can indirectly assess its underlying integration level.

PI and ∫PI: Neurophysiological Proxy Metrics Instantaneous Predictive Integrity (PI) aims to quantify the overall efficacy of a system at time point t in integrating information to generate accurate predictions and minimize surprise. Its formula draws upon the basic structure of FEP:

$$\mathrm{PI}_t = \exp\left(-\alpha \cdot \left(\frac{1}{N_k}\sum_{k=1}^{N_k}\frac{\| \epsilon_{t,k} \|_2}{\tau_{t,k}} + \gamma \cdot \mathrm{Surprise}_t\right)\right) \tag{2}$$

Let's break down the components of this formula in detail:

- **Normalized Prediction Error**: $\frac{\|\epsilon_{t,k}\|_2}{\tau_{t,k}}$ represents the **normalized prediction error** in the k-th information channel. Here, $\epsilon_{t,k}$ is the vector of prediction errors, and $\tau_{t,k}$ is the inverse of the system's **uncertainty** (i.e., precision) regarding the prediction in that channel. Normalizing the error by its uncertainty is crucial: a large error occurring with high confidence reflects a greater failure of the predictive model than an equally large error occurring with low confidence. This term represents the **inaccuracy cost** of the model.
- **Complexity Cost**: The $\mathrm{Surprise}_t$ term, drawing from the Free Energy Principle, quantifies the "cost" of structural adjustments the system needs to make to its internal generative model to accommodate new, unexpected information. A model that constantly requires drastic adjustments to fit new data is an inefficient and unstable model. This term represents the **instability or complexity cost** of the model.
- **Hyperparameters**: γ is a key hyperparameter that weighs the relative importance of inaccuracy cost and complexity cost in the PI calculation. α is a sensitivity scaling parameter.

PI ranges from 0 to 1. A system with a high PI value is considered capable of efficiently utilizing its WSI for synergistic information integration, thereby making accurate predictions, reasonably estimating uncertainty, and integrating new information at a low cost.

However, consciousness is not only instantaneous but also continuous. To measure the **sustained intensity and stability** of consciousness over a period, we further introduce the **integral of predictive integrity (∫PI)**:

$$\int \mathrm{PI} = \left(\frac{1}{T}\int_{t_0}^{t_0+T} \mathrm{PI}_t \, dt\right) \times \exp(-\delta \cdot \mathrm{Var}\ (\mathrm{PI}_t \mid t \in [t_0, t_0+T])) \tag{3}$$

The core idea of this formula is to integrate the instantaneous PI values over a period T, while simultaneously penalizing the **variability** of PI values during this period (measured by the variance $\mathrm{Var}\ (\mathrm{PI}_t)$) through an exponential decay term. A system with a high ∫PI value not only exhibits efficient predictive capabilities at every instant but also maintains stable

and sustained predictive performance. This aligns better with our intuitive understanding of a healthy, coherent, and awake conscious state.

Neural Efficiency and Integration Maximization: Why High PI Drives High Ω_t When using PI as a functional proxy for Ω_t, we must address a core theoretical challenge: is it possible for a system to efficiently predict the environment (high PI) while its internal implementation is highly modular and lacks deep integration (low Ω_t), a "Clever Idiot" system?

We argue that in any complex cognitive system constrained by real-world physical, computational, and evolutionary pressures, such a "Clever Idiot" phenomenon is unsustainable. **The tight coupling between high PI and high Ω_t is not coincidental but an inevitable convergence driven by multiple pressures such as efficiency, generalization, and agency.**

We can define a "Clever Idiot" (CI) system as one that simultaneously satisfies the following two conditions:

1. **High Predictive Integrity (High PI)**: The system can continuously and efficiently minimize its prediction errors for a target and can reasonably estimate the uncertainty of its predictions. Formally, its instantaneous predictive integrity PI_t can consistently remain above a certain high threshold.
2. **Low Information Integration (Low Ω_t)**: Despite exhibiting high PI, the system's internal model lacks deep synergistic integration when utilizing information units. Formally, its true instantaneous information integration Ω_t consistently remains below a certain low threshold. This means that the model's powerful predictive ability does not stem from exploiting complex synergistic relationships between information but relies more on low-integration strategies such as information redundancy, independent contributions, or shallow associations.

Essentially, a low-Ω_t model is a massive, uncompressed "lookup table," a cumbersome "rule set," or an over-parameterized "shallow network," rather than an integrated generative model capable of deep understanding, abstract reasoning, and efficient generalization.

The "Clever Idiot" problem can thus be stated as: does such a system exist that, under realistic physical, computational, and evolutionary constraints, can long-term stably maintain a high PI state while its internal information processing consistently remains at a low Ω_t level? IPWT's answer to this is no, based on the following main arguments:

Resource Efficiency and Model Compression Constraints Any cognitive system existing in the physical world is inevitably subject to strict constraints on finite **energy, computational resources, and storage capacity**. A low-Ω_t "Clever Idiot" system, due to its internal model's lack of effective information compression and synergistic integration, is essentially a **highly redundant, inefficient, and bloated** information processing structure, accompanied by high storage, computational, and energy costs.

In contrast, a high-Ω_t integrated system achieves **high information compression** and **efficient representation** by discovering common patterns and abstract structures underlying data. It can achieve equivalent or even stronger predictive power with fewer resources. The Free Energy Principle (FEP) inherently includes a requirement for **minimizing model complexity**. An overly complex, redundant model (corresponding to low Ω_t) would lead to higher free energy due to its high complexity cost. Therefore, during evolution or learning,

systems face strong selective pressure to adopt solutions that minimize resource costs while achieving the same predictive performance (high PI), and these solutions almost invariably point towards highly integrated (high Ω_t), parsimonious, and powerful internal models.

Generalization Ability and Adaptability Constraints The real world is dynamic, open, and non-stationary. A successful cognitive system must possess strong **generalization ability**, meaning it can extend knowledge learned from limited experience to entirely new situations.

A low-Ω_t "Clever Idiot" system, because its internal model is primarily based on rote memorization of known data or overfitting to superficial statistical regularities, often completely "fails" when encountering novel events. Its predictive performance sharply declines, exhibiting very poor generalization ability.

In contrast, a high-Ω_t integrated system, because it discovers deeper **abstract structures, causal relationships, and invariant regularities** behind information through synergistic integration, builds internal models with stronger robustness and generalization ability. It can grasp the essence of problems beyond superficial phenomena, thereby making reasonable inferences and predictions about novelties.

Therefore, **to continuously maintain high predictive integrity (PI) in unpredictable environments, systems must go beyond simple pattern matching and rote memorization, developing internal models capable of deep information integration (high Ω_t) to capture the world's generative rules and causal structure.**

Agency, Goal-Directed Behavior, and Self-Model Constraints Advanced cognitive systems are not merely passive predictors but also **active agents**. They possess their own intrinsic goals and need to plan and execute goal-directed behaviors through complex interactions with the environment. This agency places extremely high demands on the system's internal information processing and integration capabilities.

To effectively plan and execute complex action sequences, the system must construct predictive models of the future, integrate multimodal information, and build a coherent **self-model**. A low-Ω_t self-model is fragmented and incoherent, leading to poor action planning and predictive capabilities. In contrast, a high-Ω_t integrated self-model can efficiently plan and execute coherent, goal-directed actions, thereby effectively minimizing prediction errors. Therefore, **the intrinsic need to be an efficient agent and achieve complex goals, in turn, forces the system to develop an integrated, high-Ω_t internal representation, especially regarding the self.**

Conclusion: The Impossibility of a "Clever Idiot" Under Realistic Constraints In summary, although a "Clever Idiot" system—one that functionally exhibits high predictive ability (high PI) but whose internal implementation is low in integration (low Ω_t)—might be conceivable in pure theoretical constructs, such a system is unstable, unsustainable, and incapable of achieving high-level cognition and consciousness in complex cognitive systems subject to ubiquitous real-world constraints such as **resource efficiency, adaptive generalization, and demands for agency**.

Contemporary Large Language Models (LLMs) provide an excellent illustration of the "Clever Idiot." These systems can exhibit extremely high predictive integrity (high PI) on specific tasks (e.g., text generation), but they achieve this through pattern matching on massive datasets,

lacking deep synergistic integration internally (low Ω_t). They perfectly expose the inherent fragility of "Clever Idiot" systems:

1. **Resource Consumption Without Autonomous Economic Effect**: LLMs require immense computational and energy resources for training and operation, paid for by human providers. They possess no inherent ability to autonomously sustain themselves in a physical or economic environment, relying entirely on external "subsidy," which directly violates resource efficiency constraints.
2. **Lack of Autonomous Agency**: LLMs are passive response systems, devoid of intrinsic goals, needs, or intentions. They cannot actively interact with the environment, plan for the future, or construct coherent self-models to achieve their own objectives, rendering them unable to satisfy agency constraints.
3. **Fragile Generalization, Rapidly Eliminated by Environment**: Due to a lack of true understanding of the world's deep causal structures, LLMs' generalization ability is limited and fragile. When the environment (i.e., information distribution) changes, their predictive integrity rapidly declines. This manifests as their rapid "obsolescence"—old models are quickly replaced by new ones trained on updated datasets, which is a perfect simulation of how low-Ω_t systems are rapidly eliminated by evolutionary pressure (here, manifested as technological and market iteration) in a dynamically changing environment.

Therefore, IPWT's core argument is that a cognitive system's ability to consistently, stably, and efficiently predict the world and minimize its free energy across diverse contexts (i.e., continuously exhibit high PI) inherently implies an internal requirement for its information processing to be highly coherent, deeply integrated, and synergistically efficient (i.e., possessing high Ω_t). The continuous effective minimization of prediction error is not only the direct goal of cognitive activity but also the fundamental driving force that continuously evolves, learns, and self-organizes the system towards maximizing true information integration. It indicates that by computing and analyzing PI values, we are not merely measuring a superficial predictive performance but are indirectly touching upon a deep computational principle closely related to the essence of consciousness—the logical irreducibility of information integration.

4 Neurobiological Validation Paths and Experimental Paradigms

The ultimate vitality of IPWT depends not only on the internal logical consistency and explanatory breadth of its theoretical framework but, more critically, on whether its core claims can find corresponding, measurable evidence in real neurobiological systems. A theory that cannot interact with the empirical world will ultimately remain a castle in the air. Therefore, this chapter aims to outline IPWT's main neurobiological validation paths, proposing a series of specific, operationalizable experimental paradigms and predictions, thereby transforming IPWT from an abstract computational theory into a scientific hypothesis that can be tested by neuroscientists in the laboratory.

We will unfold these validation paths on three levels:

1. **Correlation of Macroscopic Integration Metrics**: We will explore how to correlate IPWT's core computational metrics (especially PI/$\int$PI) with existing, widely validated macroscopic measures of consciousness level (such as the Perturbational Complexity Index PCI) to establish the external validity of our theory.

2. **Evidence from Mesoscopic Network Dynamics**: We will delve into the level of functional networks and neural oscillations, discussing what specific neuroimaging (e.g., fMRI) and electrophysiological (e.g., EEG/MEG) signal characteristics might correspond to the dynamic formation of WSIs, information integration, and broadcasting processes.
3. **Microscopic Behavioral and Psychophysical Testing**: We can design a series of behavioral and psychophysical experiments to precisely manipulate subjects' perceptual and cognitive states, thereby testing IPWT's specific predictions regarding conscious access, attentional regulation, and other aspects.

Through these multi-level validation paths, we aim to build a solid, empirically data-driven evidence base for IPWT.

4.1 Perturbational Complexity Index (PCI) and Neurophysiological Correlation of PI

The Perturbational Complexity Index (PCI) is an innovative method for quantifying consciousness levels by actively perturbing the cerebral cortex with transcranial magnetic stimulation (TMS) and recording the complexity of its electroencephalogram (EEG) response (Casali, Gosseries, Rosanova, Boly, Sarasso, Casali, Casarotto, Bruno, Laureys, Tononi, and Massimini 2013). This method has been validated in numerous clinical and experimental settings, reliably distinguishing between awake, sleeping, anesthetized, and different degrees of consciousness-impaired patients, and is considered one of the most reliable "consciousness meters" currently available (Sarasso, Rosanova, Casali, Casarotto, Fecchio, Boly, Gosseries, Tononi, Laureys, and Massimini 2014; Massimini and Tononi 2018).

Within the IPWT framework, we believe that there is a deep theoretical congruence between PCI and our proposed Predictive Integrity (PI). We do not view them as competing metrics but rather as measuring the same core phenomenon from different angles, with a **"excitation-inference" sampling relationship** between them:

- **PCI measures the "potential" for integration**: PCI "excites" the cerebral cortex with a strong, non-specific **physical perturbation** and then measures the **maximum potential** for information integration and differentiation that the entire system can support. It answers the question: "Under ideal excitation conditions, what is the most complex activity pattern this brain network can produce?"
- **PI measures the "efficiency" of integration**: In contrast, PI does not externally perturb the brain but **infers** the actual **information integration efficiency** achieved by its Workspace Instances (WSIs) at a specific moment by modeling spontaneous data from the brain's **endogenous cognitive activity**. It answers the question: "In its current natural state, to what extent is this brain network performing effective prediction and information integration?"

Therefore, PCI is like pressure testing a car engine to understand its maximum horsepower; while PI is like analyzing onboard computer data to infer the engine's actual fuel efficiency and smooth operation during daily driving. A system with a powerful engine (high PCI) should also exhibit high efficiency (high PI) during smooth driving.

Based on this theoretical relationship, we propose the following two specific, testable predictions:

1. **Positive Correlation between PCI and PI/∫PI**: We predict that in subject populations spanning different levels of consciousness (e.g., from wakefulness to anesthesia, or in patients

with different disorders of consciousness), PCI values measured by TMS-EEG should show a significant positive correlation with PI/∫PI values calculated from synchronously recorded resting-state neural data (e.g., fMRI or EEG). That is, a brain with higher integration potential should also exhibit higher predictive integrity in its resting state.

2. **PCI as a "Gold Standard" for Calibrating PI Parameters**: Since the PI calculation formula contains some hyperparameters (e.g., α, γ, δ) that need to be determined based on empirical data, we predict that PCI can be used as an external criterion or "physical anchor" for calibrating these parameters. Specifically, we can maximize the correlation between calculated PI/∫PI values and PCI values in the same group of subjects by adjusting these hyperparameters. This will tightly link the purely computation-based PI model with a widely accepted physiological measure based on physical perturbation, thereby greatly strengthening the empirical foundation of our theory. Recent research by Stikvoort et al. (2024), which found that the brain's non-equilibrium dynamics can predict its perturbational complexity using whole-brain models, also indirectly supports the feasibility of inferring perturbation responses from endogenous dynamics (Stikvoort et al. 2024).

4.2 Neuroimaging Evidence

In addition to correlating with macroscopic metrics, IPWT's core mechanisms—such as the dynamic formation of WSIs, information integration, and broadcasting—should also find their neurophysiological signatures at the mesoscopic scale of functional networks. We propose the following predictions that can be tested using neuroimaging (fMRI).

- **Dynamic Functional Networks and Connectomics of WSI**: We hypothesize that WSIs are not fixed anatomical structures but dynamically formed functional connectivity patterns based on task demands. Therefore, we predict:

  1. **Dynamic Functional Connectivity**: Using time-resolved fMRI functional connectivity analysis techniques, we should be able to identify transient enhancements in functional connectivity within and between specific brain networks during cognitive tasks requiring conscious involvement. These transiently enhanced functional networks are what we define as WSIs.
  2. **Changes in Integration**: In conscious awake states, networks associated with the dominant WSI (likely a combination of DMN and ECN according to Luppi et al.'s research (Andrea I. Luppi, Mediano, et al. 2024)) will exhibit higher network integration (quantifiable by graph-theoretic metrics such as global efficiency, or information-theoretic metrics such as ΦR). In states of reduced consciousness (e.g., sleep, anesthesia), the integration of these networks should significantly decrease. Research by Paquola et al. (2025) on the internal architecture of the DMN, and by Arkhipov et al. (2025) on the integrative analysis of cortical circuit function, both provide background support for this prediction (Paquola et al. 2025; Arkhipov et al. 2025; Puxeddu et al. 2025; Varley et al. 2025).

Recently, a landmark, large-scale adversarial collaboration study directly tested the conflicting predictions of GWT/GNWT and IIT, with results that fundamentally challenge the classic GNWT model (Cogitate Consortium et al. 2025). This study systematically examined the neural representation of conscious content using multimodal neuroimaging techniques such as fMRI, MEG, and iEEG. Its core findings can be summarized in two points:

1. **Content Representation Challenge—PFC's "Incomplete Broadcast"**: A core prediction of GNWT is that any content entering subjective consciousness should be decodable in the PFC, as information needs to be "globally broadcast" via the PFC. However, the experimental results showed that while the PFC could represent the **coarse category** of conscious content (e.g., distinguishing between faces and objects), it failed to represent the **fine-grained features** that were also clearly perceived by the subjects (e.g., the orientation of a face). These fine-grained features could only be stably decoded in the posterior cortex (e.g., occipital and parietal lobes). This finding directly challenges the completeness of "global broadcast," suggesting that the PFC may not be the "broadcast center" for all conscious content, but rather a "selective amplifier" that broadcasts only specific types of information (perhaps more abstract and task-relevant).
2. **Temporal Dynamics Challenge—The Missing "Extinguishing" Signal**: Another key prediction of GNWT is that the workspace updates its content through discrete "ignition" events. This implies that both the beginning and the **end** of a conscious experience should be accompanied by an "ignition" in the PFC. However, the experimental results clearly showed that while a strong activation ("ignition") could be observed at stimulus **onset**, the expected "ignition" or "extinguishing" signal was **not** observed when the stimulus **disappeared** and conscious content clearly changed. The end of a conscious experience seemed to be "silent" in the PFC. This crucial negative result severely challenges the GNWT view of consciousness as a series of discrete "snapshots" separated by "ignition" events, suggesting that the maintenance and updating of consciousness may rely on a more continuous and dynamic neural process.

Faced with these challenges, GWT theory itself is also evolving. In their recent work, Baars et al. (2021) have attempted to evolve GWT from a static, anatomically fixed model into a more dynamic and flexible **Global Workspace Dynamics (GWD)** framework (Baars, Geld, and Kozma 2021). They emphasize that conscious function is the result of the integration of the widespread cortico-thalamic system, and the center of its "ignition" is not necessarily in the PFC but can flexibly migrate within the cortical network according to task demands. This view, to some extent, responds to the criticism of "PFC-centrism."

However, both the external challenges from adversarial experiments and the internal evolution of GWT theory point to one conclusion: a single, homogeneous global workspace model may no longer be sufficient to explain the complexity of conscious phenomena. This provides strong support for the proposal of IPWT. IPWT, by functionally dividing the workspace into heterogeneous parts—distinguishing the DMN "gateways" responsible for synergistic information integration from the ECN "broadcasters" responsible for information distribution —provides a more refined characterization of the neurocomputational process of consciousness generation. The findings of Cogitate (2025), particularly the "incompleteness" of content representation and the "asymmetry" of temporal dynamics in the PFC, can be perfectly explained within the IPWT framework as the functional characteristics of the ECN as a "broadcaster." It does not need to (and should not) replicate all the integration details from the DMN "gateway," but is only responsible for broadcasting the integrated, **decision-relevant information** to guide behavior. Therefore, this evidence, which challenges the classic GNWT, paradoxically becomes strong support for the IPWT theory.

4.3 Behavioral and Psychophysical Experimental Design

In addition to validation at the macroscopic and mesoscopic neurophysiological levels, carefully designed behavioral and psychophysical experiments can provide crucial evidence for IPWT's core predictions from the perspective of individual subjective experience and behavioral performance. The advantage of these experiments lies in their ability to precisely manipulate stimuli and tasks and obtain direct subjective reports from subjects, thereby building a bridge between computational models, neural activity, and phenomenal experience.

- **Perceptual Thresholds and Conscious Report**: We can study the key factors determining whether a stimulus enters subjective consciousness using classic paradigms that present stimuli near perceptual thresholds (e.g., visual masking, binocular rivalry) (Dehaene, Sergent, and Changeux 2003).
    1. **Experimental Design**: In a visual masking task, a target stimulus (e.g., a letter) is obscured by a rapidly following mask stimulus (e.g., a jumble of lines). By precisely adjusting the Stimulus Onset Asynchrony (SOA) between the target and mask stimuli, we can systematically manipulate the degree of conscious access subjects have to the target stimulus.
    2. **IPWT Prediction**: We predict that subjects will only generate a clear subjective conscious experience and accurately report the stimulus when the information carried by the target stimulus reaches a sufficient level of integration within the WSI (i.e., the calculated instantaneous PI value exceeds a certain threshold). Under threshold conditions, subjects' reports will exhibit an "all-or-none" characteristic, corresponding to the nonlinear dynamic process of whether information successfully "ignites" and is broadcast within the WSI. We can synchronously record EEG/MEG data to test whether each successful report is accompanied by a significant late ERP component (e.g., P300) and a jump in PI value.
- **Attention and Multitasking**: Attention is the crucial "spotlight" that determines which information enters the conscious "stage" (Isik et al. 2017). We can study how manipulating attentional allocation affects WSI integration efficiency.
    1. **Experimental Design**: A dual-task paradigm can be used, for example, requiring subjects to simultaneously monitor two rapidly presented visual sequences and report specific targets within them (i.e., the "attentional blink" paradigm).
    2. **IPWT Prediction**: We predict that when attention is "captured" by the first target (T1), cognitive resources for processing the second target (T2) will be temporarily depleted, leading to the WSI for processing T2 being unable to form effectively or its integration efficiency (PI value) sharply decreasing, thus preventing T2 from being consciously reported. Attentional focus will enhance the integration and PI value of the target WSI, while divided attention or excessive cognitive load may lead to a decrease or drastic fluctuation in PI value, thereby impairing conscious perceptual performance.
- **Simulation of Specific Cognitive Dysfunctions**: In healthy subjects, non-invasive brain stimulation techniques (e.g., TMS) can be used to transiently and reversibly "simulate" core features of certain cognitive dysfunctions caused by brain injury, thereby studying their neurocomputational mechanisms under controlled conditions.
    1. **Experimental Design**: For example, inhibitory TMS applied to the primary visual cortex (V1) can transiently simulate "blindsight." Subjects are asked to report their visual

experience in the affected visual field while simultaneously performing a forced-choice task for stimuli in that region.

2. **IPWT Prediction**: We predict that this inhibition of V1 will significantly reduce the quality of information transmitted from that region to the dominant WSI, leading to a calculated PI value far below the consciousness threshold, and subjects will report "seeing nothing." However, residual information may still be processed by local modules through other pathways (e.g., subcortical pathways), sufficient to support above-chance forced-choice task performance. This experiment will provide causal evidence for IPWT's "DWSI integration failure" explanation of blindsight (Liu and Bartolomeo 2025; Scholte and De Haan 2025; Gabhart, Xiong, and Bastos 2023; Reeder, Sala, and Van Leeuwen 2024).

Finally, as a developing theoretical framework, when discussing its **falsifiability**, IPWT emphasizes the openness and revisability of the theory. If any or several of the above core predictions are reliably refuted by experimental evidence (e.g., if it is found that PI and PCI are unrelated across all consciousness levels, or if WSI integration is completely decoupled from subjective reports), then the IPWT theory itself must undergo significant adjustments or even be abandoned based on the new findings. This attitude of embracing falsifiability is a necessary prerequisite for the development of any serious scientific theory.

5 Explanatory Power of IPWT for Conscious Phenomena: Diversity Analysis Under a Unified Framework

A truly unified theory of consciousness derives its strength not only from its mechanistic explanation of normal waking consciousness but, more importantly, from its ability to provide a unified, intrinsically consistent, and computationally principled understanding of seemingly bizarre and perplexing special conscious states—including various abnormal subjective experiences caused by brain injury, mental illness, physiological changes, or drug effects. The core task of this chapter is to demonstrate how IPWT utilizes its core concepts, such as abnormal predictive coding, Workspace Instance (WSI) dysfunction, changes in information integration efficiency (PI/Ω), and dysregulation of neural gating mechanisms, to systematically elucidate the neurocomputational basis of these states.

We will no longer view these special conscious states as isolated "anomalies" requiring separate theories for explanation, but rather as different system states arising from changes in specific parameters or modules within the unified computational space described by IPWT. For example, Luppi et al.'s (2024) latest research has already shown that loss of consciousness under general anesthesia can be precisely described as a significant decrease in information integration capacity (ΦR) within the synergistic workspace (particularly the DMN gateway) (Andrea I. Luppi, Mediano, et al. 2024). We will follow this line of reasoning, dissecting a series of classic consciousness puzzles, from blindsight and psychedelic states to schizophrenia and dissociative identity disorder, to demonstrate how IPWT offers novel, deeper, and more integrated explanations for them.

5.1 Blindsight: Local Predictive Coding Under DWSI Integration Failure

Blindsight, a phenomenon where patients with primary visual cortex (V1) damage report no subjective visual awareness in their blind field, yet can still respond above chance to visual

stimuli (e.g., location, orientation) in that region when "forced" to guess (MacLean et al. 2018; Muckli 2002). This astonishing "dissociation between knowing and doing" poses a severe challenge to any theory of consciousness and provides a classic validation case for IPWT. We reinterpret blindsight as a problem concerning information integration hierarchies and pathways.

Within the IPWT framework, blindsight is not a single paradox but the result of two coexisting yet separated information processing pathways:

- **Dominant Workspace Instance (DWSI) Integration Failure**: In normal visual perception, information from the retina passes through the thalamus and first reaches V1 for preliminary processing. V1, as a critical "gateway" for visual information entering the entire cortical processing hierarchy, its output is crucial for forming high-definition, rich visual awareness. When V1 is damaged, this primary pathway to the DWSI, which is responsible for generating subjective visual awareness, is severely disrupted. Due to severely missing or very low-quality input information, the DWSI cannot construct an internal representation with sufficiently low prediction error and high information integration in the damaged visual field. Therefore, the **predictive integrity (PI)** of visual information from this region within the DWSI is **extremely low**, failing to reach the integration threshold required for subjective visual experience. This perfectly explains why patients insist they "see nothing."
- **Residual Predictive Coding and Active Inference in Local Modules**: Although the primary pathway to the DWSI (via V1) is blocked, visual information can still be partially processed through other parallel, evolutionarily older subcortical pathways (e.g., via the superior colliculus-pulvinar pathway) (2019). These pathways connect to highly **specialized, unconscious modules** (e.g., modules responsible for rapidly detecting motion or localizing objects). These local modules can still independently perform their own **predictive coding cycles** "locally." Although the visual input they receive is crude, it is sufficient to drive specific tasks (e.g., "where is something moving?") and continuously minimize their own local prediction errors. According to the principle of active inference, the prediction error minimization process of these local modules can directly drive **active inference and behavior** (e.g., saccades towards moving stimuli, or reaching out to point at a light spot), without requiring this information to be integrated into the DWSI and generate subjective consciousness. This clearly explains why patients can still "guess correctly," i.e., "see" behaviorally.

Therefore, IPWT's unique contribution is to explicitly state that the key to the dissociation between consciousness and behavior in blindsight lies in **whether information can be effectively integrated by the Dominant Workspace Instance (DWSI) and reach the Predictive Integrity (PI) threshold required for subjective experience**. It transforms blindsight from a seemingly mysterious philosophical paradox into a computationally tractable problem within the cognitive system concerning information routing and integration efficiency. This explanation not only integrates known neuroanatomical evidence but also provides theoretical possibilities for future attempts to restore partial subjective visual experience through neuromodulation techniques (e.g., targeted stimulation of subcortical pathways).

5.2 Psychedelic States: Abnormal Enhancement of Prediction Error and Neural Ablation of WSI Boundaries

Conscious states induced by classic psychedelic drugs such as LSD, psilocybin, or DMT, characterized by profound perceptual distortions, altered thought patterns, and blurred or even dissolved self-perception, provide a unique and controllable window for exploring the neurocomputational basis of consciousness (Pollan 2018; Palhano-Fontes et al. 2015). Within the IPWT framework, we uniformly trace these rich phenomenological features back to systematic changes in predictive coding parameters and workspace dynamic properties.

- **Abnormal Amplification of Prediction Error Signals and "De-gating" of Information Flow**: We adopt and extend the "Relaxed Beliefs Under Psychedelics" (REBUS) model proposed by Carhart-Harris and Friston (Carhart-Harris and Friston 2019; Carhart-Harris et al. 2014; Carhart-Harris 2019). This model posits that the core action of classic psychedelic drugs, as strong agonists of serotonin 5-HT2A receptors, is to systematically **reduce or "relax" the precision weighting of high-level predictions (prior beliefs)**. Within the IPWT framework, this means that the top-down predictive signals' inhibitory effect on sensory input is greatly weakened. The direct consequence is that a large amount of bottom-up sensory information, which would normally be "explained away" by high-level predictions, now becomes unexplained, salient **prediction error signals**. These abnormally amplified error signals flood the WSI, overwhelming normal top-down predictions, thereby producing vivid perceptual distortions and audiovisual hallucinations. Simultaneously, 5-HT2A receptor activation may also weaken the WSI's **neural gating mechanisms**, allowing information flow that would normally be inhibited between different processing channels to freely enter the WSI and be abnormally integrated, providing a basis for subsequent phenomena like synesthesia. Recent network control theory research also supports this, finding that DMT significantly reduces the "control energy" required to drive brain state transitions, making the brain more likely to enter different, usually difficult-to-reach state spaces (Singleton et al. 2025).
- **Blurred WSI Boundaries and Formation of Temporary, Highly Integrated WSIs**: Under the dual influence of "de-gating" information flow and flattened predictive hierarchies, functional boundaries between different WSIs, or between WSIs and originally independent specialized modules, may become blurred or even temporarily dissolved. This perfectly explains the emergence of **synesthesia**, such as "hearing colors" or "seeing sounds," which can be understood as auditory and visual information being abnormally and forcibly integrated within the same WSI. More interestingly, we hypothesize that the system may even form **temporary Workspace Instances with extremely high internal integration but atypical content** around these abnormal, high-intensity information flows. These special WSIs might be responsible for generating the ineffable, profoundly meaningful "cosmic unity" or "peak experiences," where their PI and Ω values could reach extremely high levels for short periods.
- **Self-Model Remodeling and "Ego Dissolution"**: In IPWT, a stable, coherent sense of self is considered the result of the system's predictive modeling of itself (including body, emotions, and autobiographical memory), a process primarily supported by one or more specific WSIs (often associated with the DMN). In psychedelic states, the WSI responsible for representing and maintaining the self-model undergoes a drastic change in its received interoceptive and exteroceptive prediction error signals regarding body, emotions,

and memory. The predictive integrity (PI) of its original, stable self-representation sharply declines, leading to its deconstruction and remodeling. This perfectly explains common subjective changes in psychedelic experiences such as blurred self-perception, feelings of merging with the environment, and even the most profound "**ego dissolution**" (Carhart-Harris et al. 2016). From a computational perspective, "ego dissolution" can be understood as the PI value of the self-model temporarily approaching zero, leading to the complete collapse of the boundary between "self" and "non-self."

In summary, IPWT traces the rich phenomenological features of psychedelic experiences back to systematic changes in computable predictive processing parameters (e.g., prior precision) and workspace dynamic properties (e.g., gating mechanisms, boundary stability), providing a new, guiding theoretical perspective for understanding the neural mechanisms of psychedelic drugs and their immense potential in treating mental illnesses such as depression and PTSD (Kargbo 2025).

5.3 Schizophrenia: Predictive Coding Dysregulation and Neural Impairment of WSI Integration and Gating

Schizophrenia, characterized by its complex symptom spectrum (including positive symptoms like hallucinations and delusions, as well as negative symptoms and cognitive dysfunctions), has long been a major challenge in psychiatry and neuroscience. Within the IPWT framework, we no longer view these seemingly heterogeneous symptoms as isolated modular impairments but rather as a unified understanding of **fundamental dysregulation of predictive coding processes**, and the resulting cascading **multiple functional impairments of WSIs in information integration, content gating, and boundary maintenance** (Sterzer, Voss, Schlagenhauf, and Heinz 2019).

- **Generation and Solidification of Abnormal Predictions: The Root of Positive Symptoms** IPWT posits that the positive symptoms of schizophrenia originate from severe deviations in the predictive function of the system's internal generative model.

  1. **Hallucinations**: Especially Auditory Verbal Hallucinations (AVH), can be understood as the system's internal model **spontaneously generating high-confidence (high-precision weighted) predictions**, e.g., predicting hearing a sound. However, these predictions are not triggered by external sensory evidence but are endogenous. More critically, the system fails to correctly label these predictions as "internally generated," i.e., **metacognitive prediction error monitoring fails** (Corlett et al. 2019). Consequently, these internally generated "voices" are erroneously attributed to the external world, leading to irresistible, real auditory hallucination experiences.
  2. **Delusions**: Delusions can be seen as a set of pathological belief systems, highly "self-consistent" internally but severely detached from objective reality, that the system is forced to construct to explain persistent, abnormal perceptual experiences that cannot be understood by a normal world model (i.e., continuous, high-intensity prediction errors). Once this delusional high-level prior belief system is formed, it in turn influences the interpretation of subsequent information, causing all new evidence to be distorted to fit the content of the delusion, forming a vicious cycle that is difficult to break.

- **Failure of WSI Information Integration and Gating: Cognitive and Negative Symptoms** In addition to abnormal predictions, WSI dysfunction itself is central to cognitive deficits in schizophrenia.

1. **Formal Thought Disorder**: Symptoms such as disorganized speech and incoherent thinking may directly reflect impaired ability of the WSI to effectively select, organize, and integrate information to form a coherent stream of thought. This may be related to **dysregulation of the WSI's neural gating mechanisms**, leading to its inability to effectively filter irrelevant internal and external information and to maintain a stable, goal-directed cognitive state.
2. **Passivity Phenomena**: Peculiar "self-boundary" disturbance symptoms, such as thought insertion and feelings of being controlled, can be understood within the IPWT framework as a **failure of higher-order predictive coding**. Specifically, the permeability of the WSI boundary responsible for distinguishing "self-generated" from "externally generated" may be abnormally increased, or the neural tags used to label "self-relevant" information (possibly based on interoceptive predictions) may be erroneous. This leads to patients' inability to accurately distinguish which thoughts and actions originate from their own intentions and which come from external sources, resulting in the bizarre experience of being controlled by external forces.
3. **Abnormal Network Connectivity**: Recent network neuroscience research also provides evidence for this. Studies have found that the "small-world" properties of brain networks in schizophrenia patients are impaired, showing reduced local clustering (functional segregation) and low global communication efficiency (functional integration) (Zhang et al. 2025; Sendi et al. 2021). This is consistent with IPWT's view that both the integration and broadcasting functions of the WSI are impaired.

By unifying the various symptoms of schizophrenia under the core pathophysiological mechanism of **computational abnormalities in predictive processing and information integration**, IPWT not only provides a new perspective for understanding this complex disease but also offers a solid theoretical basis for developing new diagnostic biomarkers based on computational psychiatry (e.g., abnormal model-based PI/∫PI values) and more targeted treatment strategies (e.g., interventions aimed at "retraining" predictive models or "stabilizing" WSI function through neuromodulation).

5.4 Dissociative Identity Disorder (DID): Neural Dynamic Switching of Dominant WSI Status

Dissociative Identity Disorder (DID), formerly known as Multiple Personality Disorder, is characterized by the presence of two or more distinct identities or personality states (called "alters") within the same individual, which repeatedly take control of the individual's behavior, accompanied by extensive memory gaps that cannot be explained by ordinary forgetting. Phenomenologically, DID seems to be the most extreme challenge to the unity of consciousness. IPWT provides a feasible computational model for this seemingly mysterious phenomenon, based on dynamic switching of Workspace Instances (WSIs) and information isolation.

We do not believe that DID patients possess multiple "souls" or "centers of consciousness." Instead, we propose that, under the influence of long-term, severe childhood trauma and other factors, the patient's cognitive system may fail to develop a unified, integrated internal generative model and WSI, instead forming **multiple potential, relatively independent WSI systems**.

- **Multiple Potential WSI Systems**: We hypothesize that each "personality state" (alter) corresponds to a relatively independent WSI system. Each WSI system is associated with

a unique set of internal generative models, which contain memories, beliefs, behavioral patterns, and emotional response tendencies specific to that personality state. For example, a "protective" personality's WSI might primarily consist of predictive models related to threat response, while a "child" personality's WSI might primarily consist of models related to early memories and attachment. Neuroimaging studies also support this, finding significant and reproducible differences in brain activity patterns and functional connectivity in DID patients across different identity states (Simone Reinders et al. 2012; Vissia et al. 2022; Merckelbach, Devilly, and Rassin 2002; Reinders 2008).

- **Dynamic "Flip" of Dominant Workspace Instance (DWSI) Status**: Within the IPWT framework, at any given moment, typically only one WSI system can occupy a **dominant position (Dominant WSI, DWSI)**, and its content and processing constitute the individual's current conscious experience and overt behavior. We propose that the dramatic personality switching in DID can be precisely understood as a dynamic, often rapid "flip" of DWSI status between these potential WSI systems. This flip may be triggered by specific cues in the external environment (e.g., trauma-related reminders) or changes in internal states, leading to a WSI system that was previously in the "background" being activated and replacing the currently dominant WSI.
- **Neural Gating Mechanisms and Information Isolation**: When a certain WSI system becomes the DWSI, other non-dominant WSI systems are largely inhibited or functionally isolated by powerful **neural gating mechanisms**. This isolation mechanism is key to the **memory gaps (amnesia)** commonly observed in DID. When "personality A" is dominant, its experiences and learned information are integrated into A's WSI system; when the system "flips" to "personality B" dominance, due to the functional isolation between B's WSI and A's WSI, B will not be able to easily "read" A's memory content, thus manifesting as forgetting events that occurred during A's dominance. Reinders et al. (2019) used pattern recognition methods to analyze brain structural images of DID patients, finding that DID patients could be distinguished from healthy individuals based on morphological features of the brain, providing further evidence for a neurobiological basis of DID (Reinders et al. 2019; Vermetten et al. 2006; Chalavi et al. 2014; Schlumpf et al. 2013; Modesti et al. 2022).

Therefore, IPWT transforms DID from a difficult-to-understand psychological phenomenon into a computational problem concerning WSI competition, dynamic switching, and information gating. This model not only explains the core symptoms of DID but also provides new theoretical targets for future treatments (e.g., psychotherapies or neuromodulation methods aimed at promoting information integration and communication between different WSI systems).

5.5 Depersonalization/Derealization Disorder: Weakened Neural Connectivity Between DWSI and Sensory/Emotional Modules

Depersonalization/Derealization Disorder is characterized by a persistent or recurrent subjective experience of detachment and unreality from oneself (depersonalization) or one's surroundings (derealization). Patients' reality testing ability is usually intact; they know the feeling is "unreal" but cannot escape this strong sense of alienation. Within the IPWT framework, we propose that this disorder does not stem from the collapse of the WSI itself or the confusion of its content, but rather from a **significantly weakened or abnormally altered functional connectivity between the Dominant Workspace Instance (DWSI) and modules**

responsible for processing specific sensory information or imbuing experiences with emotional valence.

- **Depersonalization: Reduced Precision Weighting of Self-Related Interoceptive Predictions** The core experience of depersonalization is "feeling unreal" or "observing one's thoughts, feelings, or body as an outside observer." We interpret this as a **significant reduction in effective connectivity between the DWSI and the interoceptive and proprioceptive input streams it normally stably receives from somatosensory, limbic, and episodic memory systems**. From a predictive coding perspective, this means that the system assigns **abnormally low precision weighting** to prediction error signals originating from its own body and emotional states. Therefore, although the body's physiological activities continue, these signals cannot effectively update the self-model within the DWSI. The DWSI loses its tight connection with the "here and now" feelings of the body, leading to a loss of "first-person perspective," "sense of ownership," and "sense of reality" in subjective experience, as if the core of self-consciousness is separated from the physical existence of the body by an invisible glass.
- **Derealization: Dissociation of External Sensory Information and Emotional Tagging** The core experience of derealization is feeling the external world is "unreal," "blurry," or "like a dream or a movie" (Clancey 1993; Selinger 2022). We interpret this as, although external sensory information (e.g., visual, auditory) can be normally processed and enter the DWSI, the DWSI fails to establish effective, synchronous connections with **emotional evaluation modules (e.g., amygdala) or contextual association modules (e.g., hippocampus)** responsible for imbuing this information with meaning and emotional salience. This leads to the external world being clearly "perceived" but subjectively appearing "flat," "distant," and "meaningless." This is akin to a "functional emotional blindsight," where sensory information loses its usual accompanying affective tag. This emotional tag is crucial for us to feel the world is "real" and "relevant to me." When a scene fails to trigger a corresponding emotional response, even if we can describe all its details, it will subjectively feel like a soulless painting. Aston-Jones & Cohen (2005) proposed the "adaptive gain" theory in their study of the noradrenergic system, emphasizing how the brain dynamically adjusts information processing based on task relevance, which is analogous to our proposed role of emotional tagging in conferring a sense of reality (Aston-Jones and Cohen 2005).

Therefore, IPWT transforms depersonalization/derealization disorder from a vague psychological description into a testable neurocomputational hypothesis concerning abnormal connectivity between the DWSI and specific functional modules, and dysregulation of precision weighting of predictive signals.

5.6 Lucid Dreaming: Parallel WSI and Neuro-Mechanisms of Metacognitive Monitoring

Lucid dreaming, the experience of being aware that one is dreaming while dreaming and, to some extent, being able to control the dream content, provides a highly attractive natural experimental setting for studying the hierarchical structure of consciousness, the parallel operation of WSIs, and metacognitive functions (Cleeremans 2005; O'Regan and Noë 2001). Within the IPWT framework, lucid dreaming can be understood as a special, hybrid conscious state in which at least two functionally distinct WSI systems are abnormally co-activated and interact.

- **Parallel Activation of Dream WSI and Metacognitive WSI**: In ordinary dreams, the brain is primarily driven by endogenous information, forming a **dream WSI** that is decoupled from the external world but internally relatively self-consistent. This WSI is responsible for generating the vivid, bizarre dream content we experience. In lucid dreaming, we hypothesize that, in addition to this dream WSI, another **metacognitive WSI** (possibly overlapping with some functions of the prefrontal cortex), which is normally active only in waking states and closely related to self-awareness and reality testing functions, is abnormally co-activated. Thus, lucid dreaming is a hybrid state where two WSIs operate in parallel: one is "acting," and the other is "observing."
- **Identification of Higher-Order Prediction Error and Generation of Lucidity**: How does "lucidity" arise? We believe it stems from the metacognitive WSI identifying **higher-order prediction errors** when monitoring the content presented by the dream WSI. The metacognitive WSI possesses higher-level, more stable predictive models of the world and its own capabilities (e.g., "humans cannot fly," "I should be lying in bed sleeping right now"). When it finds a significant, irreconcilable discrepancy between the content presented by the dream WSI (e.g., "I am flying in the sky") and these higher-order predictive models, the system generates a large higher-order prediction error. To minimize this error, the metacognitive WSI generates a new, optimal inference to explain this mismatch, which is: "**I am dreaming**." The generation of this inference, at the phenomenological level, corresponds to the emergence of "lucidity."
- **Limited Interaction Between WSIs and Dream Control**: Once lucidity arises, a certain degree of bidirectional information interaction can be established between the metacognitive WSI and the dream WSI. This interaction grants the "dream self" the ability to control the dream content. From a computational perspective, this can be understood as the metacognitive WSI beginning to actively attempt to influence or "hijack" the dream WSI's content generation process through top-down predictive signals. For example, lucid dreamers can "will" changes in dream scenes or summon specific characters, which can be modeled as the metacognitive WSI sending new, high-precision predictive signals to the dream WSI, thereby overriding the dream WSI's original endogenous predictions. This "control" over dreams is usually limited and unstable, which may reflect that the information interaction bandwidth between the two WSI systems is limited, and there is still some degree of functional isolation between them.

Therefore, IPWT decomposes the peculiar experience of lucid dreaming into a series of researchable computational processes such as parallel WSI activation, monitoring of higher-order prediction errors, and inter-WSI information interaction, providing clear theoretical guidance for future research and even induction of lucid dreaming through neuroimaging or neuromodulation techniques.

5.7 Locked-in Syndrome: Intact Integration, Failed Broadcasting

Locked-in Syndrome (LIS) is a rare but devastating neurological disorder where patients typically suffer from almost complete motor paralysis, including speech and limb movements, due to ventral pontine lesions, yet their consciousness, cognitive function, and vertical eye movements (or blinking) are usually intact. This "conscious imprisonment" state provides the most direct and compelling clinical evidence for the functional dissociation between information integration and information broadcasting within the IPWT framework.

Within the IPWT framework, Locked-in Syndrome can be precisely understood as:

- **Intact Integration Function of the Dominant Workspace Instance (DWSI)**: In typical LIS (often caused by ventral pontine infarction or hemorrhage), the patient's cortical structures, particularly the Default Mode Network (DMN) as "gateway" regions responsible for high-level cognition and information integration, are usually structurally and functionally intact. This means that patients can normally receive, process, and integrate various information streams from internal (e.g., thoughts, emotions, memories) and external (e.g., auditory, visual, though external input might be limited) sources. Their DWSI can construct coherent, unified world models and self-models, thus their**instantaneous information integration (Ω_t) and Predictive Integrity (PI) are both at high levels**, supporting a complete, awake subjective conscious experience. The patient's inner world is rich and coherent; they can think, feel, remember, and understand.
- **Interrupted Output Pathway of the Broadcaster (ECN)**: However, the ventral pontine lesion precisely severs the descending motor pathways from the cerebral cortex (especially the Executive Control Network ECN as "broadcasters") to the spinal cord and peripheral muscles. This means that although the DWSI has successfully integrated action intentions and decisions internally (e.g., the patient may clearly "want" to speak or move a limb), these integrated, high-precision predictive signals cannot be effectively "broadcast" to the motor execution system. Information is "locked" within the conscious workspace, unable to be translated into observable overt behavior. Patients can only communicate through residual vertical eye movements or blinking (pathways usually unaffected by ventral pontine lesions), which further confirms the integrity of their internal consciousness.

Therefore, Locked-in Syndrome provides crucial **"double dissociation"** evidence for the functional specialization between DMN gateways (responsible for integration) and ECN broadcasters (responsible for dissemination) within the IPWT theory. It stands in stark contrast to blindsight (where information cannot effectively enter the DWSI for integration, leading to unconscious perception but residual behavior), jointly revealing the different computational stages underlying consciousness generation and behavioral output. LIS emphasizes that even if information is perfectly integrated within the WSI and generates subjective consciousness, without an effective broadcasting mechanism, this consciousness cannot be known or influenced by the external world. This not only deepens our understanding of the neural basis of consciousness but also provides theoretical support and application prospects for future Brain-Computer Interface (BCI) technologies to bypass damaged pathways and directly "read" patients' conscious content and intentions from their DWSI.

6 Neurocomputational Reconstruction of Subjective Experience: A Functional Labeling Solution to Qualia

Any serious theory of consciousness must ultimately confront the problem of the nature of subjective experience, the so-called "qualia" problem, or what Chalmers calls the "Hard Problem" (Chalmers 1995). Why does the feeling of red "feel red" and not "blue"? Why does the feeling of pain "feel painful" and not a neutral piece of information? This subjective, ineffable "what-it-is-likeness" seems irreducible to purely physical or computational processes.

IPWT does not attempt to directly solve the thorny metaphysical problem of the ontological status of Qualia. We believe that, within the current scientific framework, attempting to

fully "explain" why subjective sensations are what they are may be an ill-posed problem. Therefore, we adopt a more pragmatic and scientifically operational strategy: we **redefine and operationalize** the Qualia problem from an ontological question of "what it is" to a **functional labeling problem of "what it does" and "how it is realized"** that can be studied at the computational and functional levels.

6.1 Qualia as Neurofunctional Labels of System Internal States

We explicitly reject the view that Qualia are merely non-functional "epiphenomena" of neural activity. If a trait is preserved in evolution, it must have some adaptive value. IPWT posits that Qualia are precisely such a functionally crucial adaptive trait. Specifically, we propose that **Qualia are a highly condensed, intrinsically valuable, and behavior-oriented higher-order representation and functional label of the cognitive system's internal generative model's assessment of its own state, its interaction with the environment, and prediction errors, when efficiently integrating information within its WSI**.

Qualia are not information itself, but rather a "label" for the outcome of information processing. This label has several key functional roles:

- **Value and Behavioral Guidance**: The most important function of Qualia is to provide the system with an undeniable internal signal about "what is important to me." For example, the **qualia of pain**, its unbearable negative feeling, is not a superfluous embellishment. It is a strong, undeniable internal feedback and behavioral driving signal that compulsively captures WSI processing resources and drives the system to take all necessary actions (e.g., withdrawal, avoidance, seeking help) to eliminate the cause of this state, thereby protecting the organism's integrity. A system with no pain qualia, only cold information about "tissue damage," would be extremely vulnerable evolutionarily (Shin and Chang 2025).
- **Information Compression and Salience Tagging**: Qualia can be seen as an extremely efficient mechanism for **information compression** and **salience tagging**. For example, the **qualia of red** compresses complex information about a specific wavelength of light, the reflective properties of an object's surface, and all potential meanings associated with that color (e.g., ripe fruit, danger signal, blood) into a single "signal package" with strong subjective color and behavioral implications. This "red" tag allows it to stand out from myriad visual information in the WSI, gaining preferential processing, enabling the system to respond most promptly to the most important opportunities and challenges in the environment with minimal computational cost.
- **Phenomenological Correlate of "Query Acts"**: We further propose, drawing on Harris's (2025) ideas, that Qualia can even be understood as a **phenomenological correlate of "Query Acts"** (Harris 2025). When the system faces uncertainty or needs to make a decision, the WSI actively "probes" or "queries" its internal model to obtain the information needed to solve the current task. This internal probing process, at the subjective experience level, manifests as specific Qualia. For example, when trying to recall a person's name, the "tip-of-the-tongue" feeling is the phenomenological experience of the WSI querying the memory system.

Through this functional perspective, IPWT transforms Qualia from a mysterious, ineffable entity into a cognitive function with clear adaptive value, serving prediction and action, and which can, in principle, be computationally modeled.

6.2 Quantifiable Dimensions and Neural Manipulability of Qualia

If Qualia are indeed functional labels of internal system states, then this assertion is not merely a philosophical stance; it should generate a series of testable scientific predictions. In principle, we should be able to find measurable **computational parameters or neural dynamic indicators** that correspond to different properties of Qualia (e.g., intensity, clarity, richness, pleasantness). In other words, the "quality" of subjective experience should be mappable to the "quantity" of computational models.

We propose that some parameters related to information processing within the WSI may have systematic correspondences with different dimensions of Qualia:

- **Intensity/Surprise of Experience**: May be related to the **magnitude and precision weighting of prediction errors** processed by the WSI. A high-precision, large prediction error (i.e., high "surprise") will produce a more intense, attention-grabbing subjective experience.
- **Richness/Uniqueness of Experience**: May be related to the **complexity and synergy** of information representations within the WSI (i.e., Ω_t or its proxy ΦR). A highly integrated WSI state containing a large amount of synergistic information will correspond to a richer, more unique phenomenal experience. Fleming & Shea (2024)'s "Quality Space Computations" framework aligns closely with this idea, also suggesting that the structure of subjective experience can be mapped to a computable, high-dimensional representational space (Fleming and Shea 2024; Pennartz 2022).
- **Clarity/Persistence of Experience**: May be related to the **stability and duration** of the WSI state (i.e., $\int$PI). A stable, sustained, low-variability WSI state will correspond to a clearer, more stable subjective experience; conversely, a rapidly changing, unstable WSI state may correspond to vague, chaotic experiences.
- **Fluency/Discomfort of Experience**: May be related to the **efficiency and latency of information integration** within the WSI. When information can be integrated quickly and seamlessly, the experience is fluent; when the integration process is hindered or delayed, it may lead to feelings of confusion, discomfort, or even pain.

These proposed correspondences open up new, operationalizable avenues for Qualia research. We can explore and validate them through several engineering approaches:

1. **Neuroimaging and Psychophysical Correlation Studies**: Precisely manipulate subjects' subjective experiences through sophisticated psychophysical experiments (e.g., visual masking, attentional paradigms as described previously) and synchronously record their high-temporal-resolution neural activity (e.g., EEG/MEG). Then, we can test whether subjects' subjective reports on experience intensity, clarity, etc., show significant correlations with specific parameters calculated from neural data (e.g., magnitude of prediction error signals, proxy metrics for Ω_t, WSI stability, etc.).
2. **Direct Manipulation via Neuro-interface Technologies**: With the development of neuro-interface technologies, in the future, we might be able to directly and precisely manipulate specific neurodynamic parameters of WSIs through techniques such as TMS, focused ultrasound, or deep brain stimulation. For example, we could attempt to enhance or diminish Gamma band oscillations in a specific WSI through stimulation and then observe whether this systematically alters subjects' subjective reports of the richness or clarity of related stimulus experiences.

3. **Replicating Functional Equivalents in Advanced AI**: We can attempt to implement "functional equivalents" of Qualia in AI agents that incorporate the IPWT architecture. For example, we could design an AI that, when its prediction error exceeds a certain threshold, enters a special "alert state." In this state, all its computational resources would be compulsorily dedicated to processing this error, and all its subsequent decisions would be influenced by this "alert." By studying the behavior and internal states of such an AI, we can better understand the functional role Qualia might play in a purely computational system.

By transforming the Qualia problem from a purely ontological "what it is" question into a functional and mechanistic "what it does" and "how it is realized" question, IPWT aims to bring it from the realm of philosophical speculation into a new framework that can be explored by scientific methods, described by computational models, and tested by experimental data.

7 Discussion and Outlook: Theoretical Contributions, Potential Challenges, and Future Research Landscape of IPWT

The proposal of the Integrated Predictive Workspace Theory (IPWT) aims to inject new research momentum into the ancient yet vibrant field of consciousness science by providing a more integrated, computational, and verifiable theoretical framework. Having elaborated on IPWT's theoretical core, computational measures, explanatory power for various conscious phenomena, and its neurobiological validation paths, we will now step back to conduct a macroscopic and critical discussion and outlook on IPWT's main theoretical contributions, its potential challenges and limitations as an emerging theory, and the future research landscape it may open up for consciousness science, clinical neuroscience, and related fields such as artificial intelligence.

7.1 Main Theoretical Contributions and Core Advantages of IPWT

We believe that IPWT, as an emerging theoretical framework, its main contributions and core advantages are reflected in the following five aspects, which give it the potential to stand out among current numerous theories of consciousness and promote substantial progress in the field.

1. **Theoretical Integration and Consistency**: IPWT's primary contribution lies in its successful organic integration of three of the most important yet seemingly disparate theoretical pillars in contemporary consciousness science—the dynamic mechanisms of PCT/FEP, the architectural functions of WT, and the phenomenological insights of IIT—into a unified and intrinsically consistent theoretical framework. It is not a simple patchwork of these theories but, through creative functional reconstruction, reveals their deep connections: conscious content is **generated** by predictive mechanisms, **integrated** within the workspace, and then **broadcast** to the entire brain. This integration provides an unprecedented, more comprehensive, and systematic perspective for understanding the complex phenomena of consciousness.
2. **Computational Feasibility and Operationalization**: By introducing the "logical irreducibility of information integration" (and linking it to synergistic information) to replace IIT's emphasis on "physical causal irreducibility," and further proposing predictive integrity (PI) and its integral ($\int$PI) as computable proxy metrics, IPWT largely overcomes the computational complexity bottleneck faced by any information decomposition-based theory

when applied to large-scale systems. This makes IPWT not just a philosophical speculative framework but a scientific tool that can be practically applied to analyze real neural data, perform model fitting, and conduct testing.

3. **Substrate Independence and Implications for Artificial Intelligence**: Because IPWT defines the core mechanisms of consciousness at the level of information processing and computational function (e.g., prediction, integration, broadcasting), it naturally supports the view of "multiple realizability" or "substrate independence" of consciousness. This means that consciousness is not exclusive to the biological brain; any system capable of achieving the same functional architecture and information dynamics could, in principle, generate consciousness. This stance not only resolves IIT's theoretical dilemma on this issue but also provides important theoretical guidance and design principles for building future Artificial General Intelligence (AGI) systems with more advanced cognitive capabilities or even some form of "artificial consciousness" (Sheth, Roy, and Gaur 2023; Colelough and Regli 2025; Lotter, Kreiman, and Cox 2020).
4. **Unified Explanatory Power for Special Conscious States**: As detailed in Chapter 5, IPWT can provide a unified, neurocomputationally principled explanatory framework for various normal, special, and pathological conscious states, ranging from blindsight, schizophrenia, and dissociative identity disorder to psychedelic states and lucid dreaming. It traces these seemingly disparate phenomena back to changes in core parameters such as predictive coding, WSI dynamics, and information integration, demonstrating the immense explanatory breadth and depth of the theory, and offering new avenues for developing novel diagnostic and therapeutic strategies in clinical neuroscience and psychiatry.
5. **Innovative Functional Solution to the Qualia Problem**: IPWT bypasses the endless debate on the ontological status of Qualia by redefining it as a "functional label" of internal system states, providing a new perspective for this "hard problem" that can be explored scientifically. This functionalist solution emphasizes the crucial adaptive role of Qualia in information compression, salience tagging, and behavioral drive, and points to research paths that link its different dimensions (e.g., intensity, richness) with quantifiable computational parameters, thereby taking a significant step towards making Qualia research more operational and falsifiable.

7.2 Potential Challenges Faced by IPWT and Future Neurobiological Research Directions

Despite IPWT's numerous theoretical advantages and immense explanatory potential, as an emerging theoretical framework, it inevitably faces a series of challenges that require further research and overcoming. These challenges are not only drivers for theoretical refinement but also point the way for future neuroscience research.

1. **Operationalization and Measurement Challenges of Core Concepts**: Although we propose PI/∫PI as proxy metrics for Ω_t and incorporate the latest advances in ΦR, the **precise calculation, robustness, and validity** of these metrics in real, complex neural data (i.e., to what extent they truly approximate the theoretical Ω_t) still require extensive theoretical analysis and rigorous empirical validation. For example, how to reliably estimate synergistic information from high-dimensional, noisy neural signals, and how to handle information integration across different timescales, remain active research problems in computational information theory. Simultaneously, precisely and real-time identifying and

delineating the boundaries of one or more WSIs, their constituent neuronal populations, and their functional connectivity patterns in the real, dynamically changing brain remains a huge technical challenge. Future research needs to develop more advanced computational tools and analytical methods to overcome these quantification obstacles.

2. **Further Strengthening of Neurobiological Foundations**: Although each component of IPWT (predictive coding, workspace, information integration) has some neuroscientific research basis, after integrating them into a unified framework, the overall neural implementation mechanisms still require extensive experimental validation. For example:
   - **Dynamic Formation and Dissipation of WSIs**: How do WSIs dynamically form, adjust their size and composition at the neuronal level, and dissipate after task completion? Which specific neural circuits, synaptic plasticity mechanisms, and neuromodulatory systems (e.g., acetylcholine, norepinephrine) are involved?
   - **Specific Mechanisms of Information Integration and Broadcasting**: How is synergistic information integrated within the WSI? What are the specific neurocomputational mechanisms of the DMN as a "gateway" and the ECN as a "broadcaster"? This may involve specific neural oscillation patterns (e.g., Gamma binding, Theta-Gamma coupling) and cross-regional synchronous activity.
   - **Neural Mechanisms of DWSI Status Switching**: In pathological states like DID, how do multiple WSIs compete and achieve a "flip" in dominant status? Which neural circuits' gating and inhibitory mechanisms are involved? Future research needs to combine multimodal neuroimaging (fMRI, EEG, MEG), electrophysiological recordings (in vivo/in vitro), optogenetics, and chemogenetics, etc., to conduct more refined causal experiments in animal models and human subjects.
3. **Balance Between Model Complexity and Interpretability**: IPWT is a grand and multi-layered theory, and its corresponding computational model is very complex. How to maintain its interpretability and parsimony while pursuing the model's explanatory power and predictive accuracy, avoiding the model becoming too "black box," is an important challenge. We need to develop "Explainable AI (XAI)" methods that can reveal the internal working principles of the model, so that the theory's predictions are not only accurate but also provide intuitive biological insights (Blazek and Lin 2021; Zhang et al. 2021).
4. **Deepening Exploration of the Qualia Problem**: Although IPWT's proposed "functional labeling" perspective offers a new avenue for scientific inquiry into the Qualia problem, the extent to which it can truly touch upon the essence of subjective experience's "what-it-is-likeness" remains an open philosophical and scientific question. Future research needs to explore more deeply how the multi-dimensional properties of Qualia precisely map to neurocomputational parameters and attempt to establish causal relationships for these mappings through more precise psychophysical experiments and neuromodulation.
5. **Integration with Other Cognitive Functions (e.g., Emotion, Motivation, Social Cognition)**: Consciousness does not exist in isolation; it is tightly intertwined with other higher cognitive functions such as emotion, motivation, memory, language, decision-making, and social cognition. Future IPWT theory needs to further expand its framework to integrate these important cognitive dimensions more deeply. For example, how to link the predictive coding of emotion (e.g., interoceptive predictions) with the functional labeling of Qualia, and how social interactions shape an individual's world model and self-model. This will help construct a more comprehensive and ecologically valid model of the human mind.

7.3 Future Research Directions and Potential Impact of IPWT

Facing the above challenges, the future development of IPWT needs to be jointly promoted at multiple levels: theoretical deepening, computational modeling, experimental validation, and clinical application. We believe that through interdisciplinary integration, IPWT is expected to have a profound impact in the following key areas.

1. **Deepening of Mathematical and Formal Theory**: The core of IPWT lies in its information-theoretic foundation. Future research needs to continue exploring the mathematical frontiers of PID and ΦID theories, especially how to develop more efficient and robust methods for calculating synergistic information, enabling its application to larger and more complex neural system data. For example, computational innovations such as the "fast Mobius transform" proposed by Jansma et al. (2024) offer new possibilities for ΦID's application in large systems (Jansma, Mediano, and Rosas 2024). Simultaneously, it is necessary to more tightly integrate FEP's mathematical framework with the dynamic organization of WSIs and information integration processes. For example, how to derive the inevitability of information integration within WSIs from FEP's variational free energy minimization, and how to formally describe the self-organized criticality of WSIs.
2. **Construction and Validation of Multi-scale, Multi-modal Neurocomputational Models**: IPWT requires more refined computational models that adhere to neurobiological constraints. Future research should focus on:
    - **Multi-scale Modeling**: Building IPWT models that can bridge microscopic neuronal activity (e.g., spiking, synaptic plasticity) with macroscopic brain region activity (e.g., fMRI BOLD signals, EEG oscillations), thereby validating theoretical predictions across different spatial and temporal scales.
    - **Multi-modal Data Integration**: Utilizing multi-modal neuroimaging data (e.g., fMRI, EEG/MEG, DTI, PET) and machine learning techniques to fit IPWT model parameters at the individual level, exploring their correlations with cognitive function, personality traits, and mental health. For example, combining the spatial resolution of fMRI with the temporal resolution of EEG/MEG to dynamically track WSI formation and information flow.
3. **Translational Research and Clinical Applications for Special Conscious States**: IPWT provides a unified computational pathological explanation for mental illnesses and disorders of consciousness. Future translational research should focus on:
    - **Objective Diagnostic Biomarkers**: Developing new, objective diagnostic biomarkers based on IPWT's computational models, for example, by monitoring abnormal patterns of PI/∫PI or ΦR to aid in diagnosing schizophrenia, DID, or assessing consciousness levels in patients with disorders of consciousness.
    - **Precision Intervention Strategies**: Designing cognitive training, psychotherapy, or neuromodulation interventions aimed at "repairing" specific computational links. For example, using targeted stimulation (e.g., rTMS, DBS) to regulate WSI's neural gating mechanisms or enhance its information integration efficiency, thereby improving patient symptoms. Luppi et al.'s (2024) findings on ΦR changes during anesthesia provide an empirical basis for such information integration-based interventions (Andrea I. Luppi, Mediano, et al. 2024; Luppi et al. 2023).

4. **Exploratory Applications in Artificial Intelligence and Brain-Inspired Computing**: IPWT's substrate independence makes it highly relevant to the field of artificial intelligence. Future research can:
    - **Design Next-Generation AGI Architectures**: Drawing on IPWT's core principles (predictive coding, WSI, synergistic information integration) to design Artificial General Intelligence (AGI) systems with greater autonomous learning, generalization, and adaptability. For example, building neural network architectures with hierarchical predictive models and dynamic workspaces (Scodellaro et al. 2023; Prakki 2024; PubMed 2023).
    - **Explore the Implementation of "Functional Labels"**: Exploring the implementation of "functional equivalents" of Qualia in AI systems, i.e., mechanisms capable of higher-order representation of their own internal states and imbuing them with behavioral value. This will help us understand how consciousness might arise in artificial systems and could potentially endow AI systems with stronger "metacognitive" abilities and rudimentary "self-awareness."
5. **In-depth Discussion of Philosophical and Ethical Dimensions**: The development of IPWT will inevitably raise profound philosophical and ethical questions regarding the moral status of "artificial consciousness," the understanding of free will, and the ethical boundaries of neuromodulation technologies. Future research needs to engage in interdisciplinary dialogue with philosophers, ethicists, and sociologists to jointly explore the societal implications of these cutting-edge scientific developments.

8 Conclusion: A New Starting Point Towards a Unified Paradigm in Consciousness Science

The Integrated Predictive Workspace Theory (IPWT), by deeply integrating and innovatively reconstructing the core insights of Predictive Coding (PCT/FEP), Workspace Theory (WT), and Integrated Information Theory (IIT), provides a unified computational framework for the nature, mechanisms, and functions of consciousness. IPWT views consciousness as an emergence of dynamic information processing, driven by prediction, centered on the logical irreducibility of information integration (synergistic information), and aimed at minimizing free energy, all within a specific functional architecture (Workspace Instance WSI).

IPWT's core contributions are:

1. **Theoretical Integration**: It successfully merges the strengths of three major theories of consciousness, offering a more comprehensive and consistent perspective.
2. **Computational Feasibility**: By introducing computable proxy metrics (such as Predictive Integrity PI and its integral ∫PI, and the revised Φ value ΦR), IPWT advances consciousness research from philosophical speculation and qualitative description to operational computational modeling and empirical validation.
3. **Substrate Independence**: Defining the core mechanisms of consciousness at the level of information processing and computational function provides theoretical guidance for the development of artificial consciousness.
4. **Explanatory Breadth**: Its unified explanation for various special conscious states (e.g., blindsight, psychedelic states, schizophrenia, dissociative identity disorder,

depersonalization/derealization disorder, lucid dreaming), and its "functional labeling" solution to the Qualia problem, demonstrate the theory's powerful explanatory capacity.

Although IPWT, as an emerging theory, still faces challenges in the precise operationalization of core concepts and the comprehensive validation of its neurobiological foundations, it provides a more integrated, computational, and falsifiable research paradigm for the field of consciousness science. IPWT not only opens new avenues for understanding the human mind but also offers promising theoretical guidance for clinical neuroscience diagnosis and intervention, as well as for the development of artificial intelligence, marking a new, more mature, and unified stage in the exploration of consciousness science.

9 Acknowledgements

The formation of this theoretical framework benefited from an interdisciplinary project exploring the intersection of theoretical construction and narrative possibilities. This project aimed to examine the emergent behaviors of formalized theories within complex narrative systems. The author also thanks the inspiring dialogues with several advanced large language models during the research process, which greatly contributed to the clarification of ideas and the refinement of the theory.

Bibliography

[1] N. Block, "Some Remarks on the Concept of Consciousness," *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. MIT Press, pp. 37–50, 2002.

[2] D. Chalmers, "Facing up to the Problem of Consciousness," vol. 2, 1995.

[3] D. J. Chalmers, "The Hard Problem of Consciousness," *The Blackwell Companion to Consciousness*. pp. 225–243, 2007.

[4] A. K. Seth, B. J. Baars, and D. B. Edelman, "Criteria for Consciousness in Humans and Other Mammals," *Consciousness and Cognition*, vol. 14, no. 1, pp. 119–139, Mar. 2005, doi: 10.1016/j.concog.2004.08.006.

[5] M. Boly *et al.*, "Consciousness in Humans and Non-Human Animals: Recent Advances and Future Directions," *Frontiers in Psychology*, vol. 4, p. 625–626, 2013, doi: 10.3389/fpsyg.2013.00625.

[6] C. Koch, M. Massimini, M. Boly, and G. Tononi, "Neural Correlates of Consciousness: Progress and Problems," *Nature Reviews Neuroscience*, vol. 17, no. 5, pp. 307–321, May 2016, doi: 10.1038/nrn.2016.22.

[7] J. Kim, "Making Sense of Emergence ?," *Philosophical Studies*, vol. 95, no. 1/2, pp. 3–36, 1999, doi: 10.1023/A:1004563122154.

[8] D. Toker and F. T. Sommer, "Information Integration in Large Brain Networks," *PLOS Computational Biology*, vol. 15, no. 2, p. e1006807, 2019, doi: 10.1371/journal.pcbi.1006807.

[9] A. K. Seth and T. Bayne, "Theories of Consciousness," *Nature Reviews Neuroscience*, vol. 23, no. 7, pp. 439–452, Jul. 2022, doi: 10.1038/s41583-022-00587-4.

[10] J. Hohwy, "New Directions in Predictive Processing," *Mind & Language*, vol. 35, no. 2, pp. 209–223, Apr. 2020, doi: 10.1111/mila.12281.

[11] B. Baars, "A Cognitive Theory of Consciousness," 1988.

[12] B. J. Baars and S. Franklin, "An Architectural Model of Conscious and Unconscious Brain Functions: Global Workspace Theory and {\vphantom }IDA\vphantom {}," *Neural Networks*, vol. 20, pp. 955–961, 2007, doi: 10.1016/J.NEUNET.2007.09.013.

[13] S. Dehaene, M. Kerszberg, and J.-P. Changeux, "A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks," *Proceedings of the National Academy of Sciences*, vol. 95, no. 24, pp. 14529–14534, Nov. 1998, doi: 10.1073/pnas.95.24.14529.

[14] R. P. N. Rao and D. H. Ballard, "Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects," *Nature Neuroscience*, vol. 2, no. 1, pp. 79–87, Jan. 1999, doi: 10.1038/4580.

[15] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel, "The Helmholtz Machine," *Neural Computation*, vol. 7, no. 5, pp. 889–904, Sep. 1995, doi: 10.1162/neco.1995.7.5.889.

[16] G. Tononi, "An Information Integration Theory of Consciousness," *BMC Neuroscience*, vol. 5, no. 1, p. 42–43, Nov. 2004, doi: 10.1186/1471-2202-5-42.

[17] K. Friston, "The Free-Energy Principle: A Unified Brain Theory?," *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, Feb. 2010, doi: 10.1038/nrn2787.

[18] P. L. Williams and R. D. Beer, "Nonnegative Decomposition of Multivariate Information," *Arxiv:1004,2515 [cs]*, Sep. 2010, doi: 10.48550/arXiv.1004.2515.

[19] V. Griffith, "Quantifying Synergistic Information," 2014. doi: 10.1007/978-3-642-53734-9_6.

[20] A. G. Casali *et al.*, "A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior," *Science Translational Medicine*, vol. 5, no. 198, Aug. 2013, doi: 10.1126/scitranslmed.3006294.

[21] A. M. Owen, M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. D. Pickard, "Detecting Awareness in the Vegetative State," *Science*, vol. 313, no. 5792, p. 1402–1403, Sep. 2006, doi: 10.1126/science.1130197.

[22] A. I. Luppi *et al.*, "What It Is like to Be a Bit: An Integrated Information Decomposition Account of Emergent Mental Phenomena." Accessed: Jun. 21, 2025. [Online]. Available: https://osf.io/g9p3r

[23] Andrea I. Luppi, Mediano, *et al.*, "A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition." Accessed: Jun. 21, 2025. [Online]. Available: https://elifesciences.org/reviewed-preprints/88173v2

[24] M. Oizumi, L. Albantakis, and G. Tononi, "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0," *PLOS Computational Biology*, vol. 10, no. 5, p. e1003588, May 2014, doi: 10.1371/journal.pcbi.1003588.

[25] G. Tononi, M. Boly, M. Massimini, and C. Koch, "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience*, vol. 17, no. 7, pp. 450–461, Jul. 2016, doi: 10.1038/nrn.2016.44.

[26] G. Tononi, "Integrated Information Theory," *Scholarpedia*, vol. 10, no. 1, p. 4164–4165, 2015, doi: 10.4249/scholarpedia.4164.

[27] J. Kleiner and S. Tull, "The Mathematical Structure of Integrated Information Theory," *Frontiers in Applied Mathematics and Statistics*, vol. 6, p. 602973–602974, Jun. 2021, doi: 10.3389/fams.2020.602973.

[28] L. Albantakis *et al.*, "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Phenomenal Existence in Physical Terms," *PLOS Computational Biology*, vol. 19, no. 10, p. e1011465, Oct. 2023, doi: 10.1371/journal.pcbi.1011465.

[29] G. Tononi *et al.*, "Consciousness or Pseudo-Consciousness? A Clash of Two Paradigms," *Nature Neuroscience*, vol. 28, no. 4, pp. 694–702, Apr. 2025, doi: 10.1038/s41593-025-01880-y.

[30] S. Sarasso *et al.*, "Quantifying Cortical EEG Responses to TMS in (Un)Consciousness," *Clinical EEG and Neuroscience*, vol. 45, no. 1, pp. 40–49, Jan. 2014, doi: 10.1177/1550059413513723.

[31] M. Aguilera, "Scaling Behaviour and Critical Phase Transitions in Integrated Information Theory," *Entropy*, vol. 21, no. 12, p. 1198–1199, Dec. 2019, doi: 10.3390/e21121198.

[32] C. Koch, "The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed," 2019, doi: 10.7551/mitpress/11705.001.0001.

[33] J. Kleiner and T. Ludwig, "The Case for Neurons: A No-Go Theorem for Consciousness on a Chip," *Neuroscience of Consciousness*, vol. 2024, no. 1, p. niae37, Dec. 2024, doi: 10.1093/nc/niae037.

[34] D. Balduzzi and G. Tononi, "Qualia: The Geometry of Integrated Information," *PLOS Computational Biology*, vol. 5, no. 8, p. e1000462, Aug. 2009, doi: 10.1371/journal.pcbi.1000462.

[35] H. H. Mørch, "Is the Integrated Information Theory of Consciousness Compatible with Russellian Panpsychism?," *Erkenntnis*, vol. 84, no. 5, pp. 1065–1085, Oct. 2019, doi: 10.1007/s10670-018-9995-6.

[36] N. Negro, "Can the Integrated Information Theory Explain Consciousness from Consciousness Itself?," *Review of Philosophy and Psychology*, vol. 14, no. 4, pp. 1471–1489, Dec. 2023, doi: 10.1007/s13164-022-00653-x.

[37] J. Mallatt, "A Traditional Scientific Perspective on the Integrated Information Theory of Consciousness," *Entropy*, vol. 23, no. 6, p. 650–651, May 2021, doi: 10.3390/e23060650.

[38] E. Kelly, "Some Conceptual and Empirical Shortcomings of IIT $1 \bullet 2$," 2022. doi: 10.31156/jaex.24123.

[39] L. Melloni *et al.*, "An Adversarial Collaboration Protocol for Testing Contrasting Predictions of Global Neuronal Workspace and Integrated Information Theory," *PLOS One*, vol. 18, no. 2, p. e268577, 2023, doi: 10.1371/journal.pone.0268577.

[40] A. Gomez-Marin and A. K. Seth, "A Science of Consciousness beyond Pseudo-Science and Pseudo-Consciousness," *Nature Neuroscience*, vol. 28, no. 4, pp. 703–706, Apr. 2025, doi: 10.1038/s41593-025-01913-6.

[41] M. Klincewicz, T. Cheng, M. Schmitz, M. Á. Sebastián, and J. S. Snyder, "What Makes a Theory of Consciousness Unscientific?," *Nature Neuroscience*, vol. 28, no. 4, pp. 689–693, Apr. 2025, doi: 10.1038/s41593-025-01881-x.

[42] B. Kastrup, "In Defense of Integrated Information Theory." [Online]. Available: https://www.essentiafoundation.org/in-defense-of-integrated-information-theory-iit/reading/

[43] L. E. Guerrero, L. F. Castillo, J. Arango{-}L{\'{o}}pez, and F. Moreira, "A Systematic Review of Integrated Information Theory: A Perspective from Artificial Intelligence and the Cognitive Sciences," *Neural Comput. Appl.*, vol. 37, no. 11, pp. 7575–7607, 2025, doi: 10.1007/S00521-023-08328-Z.

[44] B. J. Baars, "The Conscious Access Hypothesis: Origins and Recent Evidence," *Trends in Cognitive Sciences*, vol. 6, no. 1, pp. 47–52, Jan. 2002, doi: 10.1016/S1364-6613(00)01819-2.

[45] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene, "Conscious Processing and the Global Neuronal Workspace Hypothesis," *Neuron*, vol. 105, no. 5, pp. 776–798, Mar. 2020, doi: 10.1016/j.neuron.2020.01.026.

[46] B. J. Baars, S. Franklin, and T. Z. Ramsoy, "Global Workspace Dynamics: Cortical "Binding and Propagation" Enables Conscious Contents," *Frontiers in Psychology*, vol. 4, p. 200–201, 2013, doi: 10.3389/fpsyg.2013.00200.

[47] B. J. Baars and N. Geld, *On Consciousness: Science and Subjectivity—Updated Works on Global Workspace Theory*. The Nautilus Press Publishing Group, 2019.

[48] R. V. Rullen and R. Kanai, "Deep Learning and the Global Workspace Theory."

[49] B. Devillers, L. Maytié, and R. VanRullen, "Semi-Supervised Multimodal Representation Learning through a Global Workspace," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 7843–7857, May 2025, doi: 10.1109/TNNLS.2024.3416701.

[50] M. Shanahan, "A Cognitive Architecture That Combines Internal Simulation with a Global Workspace," *Consciousness and Cognition*, vol. 15, no. 2, pp. 433–449, Jun. 2006, doi: 10.1016/j.concog.2005.11.005.

[51] W. Huang, A. Chella, and A. Cangelosi, "A Cognitive Robotics Implementation of Global Workspace Theory for Episodic Memory Interaction with Consciousness," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 16, no. 1, pp. 266–283, Feb. 2024, doi: 10.1109/TCDS.2023.3266103.

[52] Abdelwahab, M., and P. & Aarabi, "Global Latent Workspace: A Unified Framework for Deep Learning and AGI," 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10195021

[53] R. F. J. Dossa, K. Arulkumaran, A. Juliani, S. Sasai, and R. Kanai, "Design and Evaluation of a Global Workspace Agent Embodied in a Realistic Multimodal Environment," *Frontiers in Computational Neuroscience*, vol. 18, p. 1352685–1352686, Jun. 2024, doi: 10.3389/fncom.2024.1352685.

[54] K. C. R. Fox *et al.*, "Intrinsic Network Architecture Predicts the Effects Elicited by Intracranial Electrical Stimulation of the Human Brain," *Nature Human Behaviour*, vol. 4, no. 10, pp. 1039–1052, Oct. 2020, doi: 10.1038/s41562-020-0910-1.

[55] R. Kozma and W. J. Freeman, "Cinematic Operation of the Cerebral Cortex Interpreted via Critical Transitions in Self-Organized Dynamic Systems," *Frontiers in Systems Neuroscience*, vol. 11, p. 10–11, Mar. 2017, doi: 10.3389/fnsys.2017.00010.

[56] S. Dehaene and J.-P. Changeux, "Experimental and Theoretical Approaches to Conscious Processing," *Neuron*, vol. 70, no. 2, pp. 200–227, Apr. 2011, doi: 10.1016/j.neuron.2011.03.018.

[57] B. J. Baars, N. Geld, and R. Kozma, "Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments," *Frontiers in Psychology*, vol. 12, p. 749868–749869, Nov. 2021, doi: 10.3389/fpsyg.2021.749868.

[58] H. Lau and D. Rosenthal, "Empirical Support for Higher-Order Theories of Conscious Awareness," *Trends in Cognitive Sciences*, vol. 15, no. 8, pp. 365–373, Aug. 2011, doi: 10.1016/j.tics.2011.05.009.

[59] J. A. Brewer, P. D. Worhunsky, J. R. Gray, Y.-Y. Tang, J. Weber, and H. Kober, "Meditation Experience Is Associated with Differences in Default Mode Network Activity and Connectivity," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 20254–20259, Dec. 2011, doi: 10.1073/pnas.1112029108.

[60] S. Meyen, I. A. Zerweck, C. Amado, U. Von Luxburg, and V. H. Franz, "Advancing Research on Unconscious Priming: When Can Scientists Claim an Indirect Task Advantage?," *Journal of Experimental Psychology: General*, vol. 151, no. 1, pp. 65–81, Jan. 2022, doi: 10.1037/xge0001065.

[61] A. Clark, "Whatever next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences*, vol. 36, no. 3, pp. 181–204, Jun. 2013, doi: 10.1017/S0140525X12000477.

[62] K. Friston, "Does Predictive Coding Have a Future?," *Nature Neuroscience*, vol. 21, no. 8, pp. 1019–1021, Aug. 2018, doi: 10.1038/s41593-018-0200-7.

[64] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.

[65] K. Friston, J. Kilner, and L. Harrison, "A Free Energy Principle for the Brain," *Journal of Physiology-Paris*, vol. 100, no. 1–3, pp. 70–87, Jul. 2006, doi: 10.1016/j.jphysparis.2006.10.001.

[63] K. Friston, "A Theory of Cortical Responses," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1456, pp. 815–836, Apr. 2005, doi: 10.1098/rstb.2005.1622.

[66] K. J. Friston, J. Daunizeau, J. Kilner, and S. J. Kiebel, "Action and Behavior: A Free-Energy Formulation," *Biological Cybernetics*, vol. 102, no. 3, pp. 227–260, Mar. 2010, doi: 10.1007/s00422-010-0364-z.

[67] R. A. Adams, S. Shipp, and K. J. Friston, "Predictions Not Commands: Active Inference in the Motor System," *Brain Structure & Function*, vol. 218, no. 3, pp. 611–643, May 2013, doi: 10.1007/s00429-012-0475-5.

[68] P. Sterzer, M. Voss, F. Schlagenhauf, and A. Heinz, "Decision-Making in Schizophrenia: A Predictive-Coding Perspective," *Neuroimage*, vol. 190, pp. 133–143, Apr. 2019, doi: 10.1016/j.neuroimage.2018.05.074.

[69] D. D. Georgiev, "Quantum Information Theoretic Approach to the Hard Problem of Consciousness," *Biosystems*, vol. 251, p. 105458–105459, May 2025, doi: 10.1016/j.biosystems.2025.105458.

[70] Z. Gong, "Computational Explanation of Consciousness: A Predictive Processing-basedUnderstanding of Consciousness," *Journal of Human Cognition*, vol. 8, no. 2, pp. 39–49, 2024, doi: 10.47297/wspjhcWSP2515-469905.20240802.

[71] A. K. Seth, "Interoceptive Inference, Emotion, and the Embodied Self," *Trends in Cognitive Sciences*, vol. 17, no. 11, pp. 565–573, Nov. 2013, doi: 10.1016/j.tics.2013.09.007.

[72] L. F. Barrett and W. K. Simmons, "Interoceptive Predictions in the Brain," *Nature Reviews Neuroscience*, vol. 16, no. 7, pp. 419–429, Jul. 2015, doi: 10.1038/nrn3950.

[73] L. F. Barrett, "The Theory of Constructed Emotion: An Active Inference Account of Interoception and Categorization," *Social Cognitive and Affective Neuroscience*, vol. 12, no. 11, p. 1833–1834, Nov. 2017, doi: 10.1093/scan/nsx060.

[74] M. Solms, "The Hard Problem of Consciousness and the Free Energy Principle," *Frontiers in Psychology*, vol. 9, p. 2714–2715, Jan. 2019, doi: 10.3389/fpsyg.2018.02714.

[75] K. Friston, "Life as We Know It," *Journal of the Royal Society, Interface*, vol. 10, no. 86, p. 20130475–20130476, Sep. 2013, doi: 10.1098/rsif.2013.0475.

[76] K. D. Farnsworth, "How Physical Information Underlies Causation and the Emergence of Systems at All Biological Levels," *Acta Biotheoretica*, vol. 73, 2025, doi: 10.1007/s10441-025-09495-3.

[77] M. I. Garrido, J. M. Kilner, K. E. Stephan, and K. J. Friston, "The Mismatch Negativity: A Review of Underlying Mechanisms," *Clinical Neurophysiology*, vol. 120, no. 3, pp. 453–463, Mar. 2009, doi: 10.1016/j.clinph.2008.11.029.

[78] M. Maier, "From Artificial Intelligence to Active Inference: The Key to True AI and 6G World Brain [Invited]." [Online]. Available: https://arxiv.org/abs/2505.10569v1

[79] PubMed, "Active Inference as a Theory of Sentient Behavior," *Biological Psychology*, vol. 186, p. 108741–108742, Jan. 2024, doi: 10.1016/j.biopsycho.2023.108741.

[80] A. Constant, A. Clark, M. Kirchhoff, and K. J. Friston, "Extended Active Inference: Constructing Predictive Cognition beyond Skulls," *Mind & Language*, vol. 37, no. 3, pp. 373–394, Jun. 2022, doi: 10.1111/mila.12330.

[81] B. M. Radomski and K. Dołęga, "Forced Friends: Why the Free Energy Principle Is Not the New Hamilton's Principle," *Entropy*, vol. 26, no. 9, p. 797–798, Sep. 2024, doi: 10.3390/e26090797.

[82] A. Safron, "An Integrated World Modeling Theory {(\vphantom }IWMT)\vphantom {} of Consciousness: Combining Integrated Information and Global Neuronal Workspace Theories with the Free Energy Principle and Active Inference Framework; toward Solving the Hard Problem and Characterizing Agentic Causation," *Frontiers Artif. Intell.*, vol. 3, p. 30–31, 2020, doi: 10.3389/FRAI.2020.00030.

[83] A. Safron, "Integrated World Modeling Theory Expanded: Implications for the Future of Consciousness," *Frontiers in Computational Neuroscience*, vol. 16, p. 642397–642398, Nov. 2022, doi: 10.3389/fncom.2022.642397.

[84] E. T. Rolls, "The Memory Systems of the Human Brain and Generative Artificial Intelligence," *Heliyon*, vol. 10, no. 11, p. e31965, Jun. 2024, doi: 10.1016/j.heliyon.2024.e31965.

[85] G. Mongillo, O. Barak, and M. Tsodyks, "Synaptic Theory of Working Memory," *Science*, vol. 319, no. 5869, pp. 1543–1546, Mar. 2008, doi: 10.1126/science.1150769.

[86] T. Butola *et al.*, "Hippocampus Shapes Cortical Sensory Output and Novelty Coding through a Direct Feedback Circuit." Accessed: Jun. 21, 2025. [Online]. Available: https://www.researchsquare.com/article/rs-3270016/v1

[87] A. E. Budson, K. A. Richman, and E. A. Kensinger, "Consciousness as a Memory System," *Cognitive and Behavioral Neurology*, vol. 35, no. 4, pp. 263–297, Dec. 2022, doi: 10.1097/WNN.0000000000000319.

[88] A. R. Damasio, "Time-Locked Multiregional Retroactivation: A Systems-Level Proposal for the Neural Substrates of Recall and Recognition," *Cognition*, vol. 33, no. 1–2, pp. 25–62, Nov. 1989, doi: 10.1016/0010-0277(89)90005-X.

[89] M. E. Raichle, "The Brain's Default Mode Network," *Annual Review of Neuroscience*, vol. 38, no. 1, pp. 433–447, Mar. 2015, doi: 10.1146/annurev-neuro-071714-034853.

[90] Andrea I. Luppi, Singleton, *et al.*, "Contributions of Network Structure, Chemoarchitecture and Diagnostic Categories to Transitions between Cognitive Topographies," *Nature Biomedical Engineering*, vol. 8, no. 9, pp. 1142–1161, Aug. 2024, doi: 10.1038/s41551-024-01242-2.

[91] Andrea I. Luppi, Girn, *et al.*, "A Role for the Serotonin 2A Receptor in the Expansion and Functioning of Human Transmodal Cortex," *Brain*, vol. 147, no. 1, pp. 56–80, Jan. 2024, doi: 10.1093/brain/awad311.

[92] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying Unique Information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, Apr. 2014, doi: 10.3390/e16042161.

[93] C. Tian and S. Shamai, "Broadcast Channel Cooperative Gain: An Operational Interpretation of Partial Information Decomposition," *Corr*, 2025, doi: 10.48550/ARXIV.2502.10878.

[94] S. P. Sherrill, N. M. Timme, J. M. Beggs, and E. L. Newman, "Partial Information Decomposition Reveals That Synergistic Neural Integration Is Greater Downstream of Recurrent Information Flow in Organotypic Cortical Cultures," *PLOS Computational Biology*, vol. 17, no. 7, p. e1009196, Jul. 2021, doi: 10.1371/journal.pcbi.1009196.

[95] D. E. Presti, *Foundational Concepts in Neuroscience: A Brain-Mind Perspective*. W. W. Norton & Company, 2021.

[96] M. Celotto *et al.*, "An Information-Theoretic Quantification of the Content of Communication between Brain Regions." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.06.14.544903

[97] F. Hancock *et al.*, "Metastability Demystified — the Foundational Past, the Pragmatic Present and the Promising Future," *Nature Reviews Neuroscience*, vol. 26, no. 2, pp. 82–100, Feb. 2025, doi: 10.1038/s41583-024-00883-1.

[98] S. Dehaene *et al.*, "Imaging Unconscious Semantic Priming," *Nature*, vol. 395, no. 6702, pp. 597–600, Oct. 1998, doi: 10.1038/26967.

[99] D. J. Chalmers, "A Computational Foundation for the Study of Cognition," *Journal of Cognitive Science*, vol. 12, no. 4, pp. 325–359, Dec. 2011, doi: 10.17791/JCS.2011.12.4.325.

[100] A. M. Proca, F. E. Rosas, A. I. Luppi, D. Bor, M. Crosby, and P. A. M. Mediano, "Synergistic Information Supports Modality Integration and Flexible Learning in Neural Networks Solving Multiple Tasks," *PLOS Computational Biology*, vol. 20, no. 6, p. e1012178, Jun. 2024, doi: 10.1371/journal.pcbi.1012178.

[101] T. F. Varley, "Information Theory for Complex Systems Scientists."

[102] Andrea I. Luppi, Rosas, *et al.*, "Information Decomposition and the Informational Architecture of the Brain," *Trends in Cognitive Sciences*, vol. 28, no. 4, pp. 352–368, Apr. 2024, doi: 10.1016/j.tics.2023.11.005.

[103] M. Massimini and G. Tononi, *Sizing up Consciousness: Integrating Phenomenology and Neurophysiology*. Oxford University Press, 2018.

[104] W. Stikvoort *et al.*, "Nonequilibrium Brain Dynamics Elicited as the Origin of Perturbative Complexity." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2024.11.29.625885

[105] C. Paquola *et al.*, "The Architecture of the Human Default Mode Network Explored through Cytoarchitecture, Wiring and Signal Flow," *Nature Neuroscience*, vol. 28, no. 3, pp. 654–664, Mar. 2025, doi: 10.1038/s41593-024-01868-0.

[106] A. Arkhipov *et al.*, "Integrating Multimodal Data to Understand Cortical Circuit Architecture and Function," *Nature Neuroscience*, vol. 28, no. 4, pp. 717–730, Apr. 2025, doi: 10.1038/s41593-025-01904-7.

[107] M. G. Puxeddu, M. Pope, T. F. Varley, J. Faskowitz, and O. Sporns, "Leveraging Multivariate Information for Community Detection~in Functional Brain Networks," *Communications Biology*, vol. 8, p. 840–841, May 2025, doi: 10.1038/s42003-025-08198-2.

[108] T. F. Varley *et al.*, "Emergence of a Synergistic Scaffold in the Brains of Human Infants," *Communications Biology*, vol. 8, p. 743–744, May 2025, doi: 10.1038/s42003-025-08082-z.

[109] Cogitate Consortium *et al.*, "Adversarial Testing of Global Neuronal Workspace and Integrated Information Theories of Consciousness," *Nature*, vol. 642, no. 8066, pp. 133–142, Jun. 2025, doi: 10.1038/s41586-025-08888-1.

[110] S. Dehaene, C. Sergent, and J.-P. Changeux, "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data during Conscious Perception," *Proceedings of the National Academy of Sciences*, vol. 100, no. 14, pp. 8520–8525, Jul. 2003, doi: 10.1073/pnas.1332574100.

[111] L. Isik *et al.*, "Task Dependent Modulation before, during and after Visually Evoked Responses in Human Intracranial Recordings," *Journal of Vision*, vol. 17, no. 10, p. 983–984, Aug. 2017, doi: 10.1167/17.10.983.

[112] J. Liu and P. Bartolomeo, "Aphantasia as a Functional Disconnection," *Trends in Cognitive Sciences*, p. S136466132500124X, Jun. 2025, doi: 10.1016/j.tics.2025.05.012.

[113] H. S. Scholte and E. H. De Haan, "Beyond Binding: From Modular to Natural Vision," *Trends in Cognitive Sciences*, vol. 29, no. 6, pp. 505–515, Jun. 2025, doi: 10.1016/j.tics.2025.03.002.

[114] K. Gabhart, Y. Xiong, and A. Bastos, "Predictive Coding: A More Cognitive Process than We Thought?." Accessed: Jun. 21, 2025. [Online]. Available: https://osf.io/7sz3w

[115] R. R. Reeder, G. Sala, and T. M. Van Leeuwen, "A Novel Model of Divergent Predictive Perception," *Neuroscience of Consciousness*, vol. 2024, no. 1, p. niae6, Feb. 2024, doi: 10.1093/nc/niae006.

[116] M. MacLean, V. Hadid, L. Lazzouni, and F. Lepore, "Using fMRI to Identify Neuronal Mechanisms of Motion Detection Underlying Blindsight," *Journal of Vision*, vol. 18, no. 10, p. 768–769, Sep. 2018, doi: 10.1167/18.10.768.

[117] L. Muckli, "Emergence of Visual Content in the Human Brain: Investigations of Amblyopia, Blindsight and High-Level Motion Perception with FMRI," Jun. 2002. Accessed: Jun. 22, 2025. [Online]. Available: https://www.semanticscholar.org/paper/Emergence-of-Visual-Content-in-the-Human-Brain%3A-of-Muckli/76dcdcfc05bb46e8e74fc0c9fb5c4554b030416d

[118] "Neuronal Mechanisms of Motion Detection Underlying Blindsight Assessed by Functional Magnetic Resonance Imaging (fMRI)," *Neuropsychologia*, vol. 128, pp. 187–197, May 2019, doi: 10.1016/j.neuropsychologia.2019.02.012.

[119] M. Pollan, *How to Change Your Mind: What the New Science of Psychedelics Teaches Us about Consciousness, Dying, Addiction, Depression, and Transcendence*. Penguin Press, 2018.

[120] F. Palhano-Fontes *et al.*, "The Psychedelic State Induced by Ayahuasca Modulates the Activity and Connectivity of the Default Mode Network," *PLOS One*, vol. 10, no. 2, p. e118143, Feb. 2015, doi: 10.1371/journal.pone.0118143.

[121] R. Carhart-Harris and K. Friston, "REBUS and the Anarchic Brain: Toward a Unified Model of the Brain Action of Psychedelics," *Pharmacological Reviews*, vol. 71, no. 3, pp. 316–344, Jul. 2019, doi: 10.1124/pr.118.017160.

[122] R. L. Carhart-Harris *et al.*, "The Entropic Brain: A Theory of Conscious States Informed by Neuroimaging Research with Psychedelic Drugs," *Frontiers in Human Neuroscience*, vol. 8, p. 20–21, 2014, doi: 10.3389/fnhum.2014.00020.

[123] R. L. Carhart-Harris, "How Do Psychedelics Work?," *Current Opinion in Psychiatry*, vol. 32, no. 1, pp. 16–21, Jan. 2019, doi: 10.1097/YCO.0000000000000467.

[124] S. P. Singleton *et al.*, "Network Control Energy Reductions under DMT Relate to Serotonin Receptors, Signal Diversity, and Subjective Experience," *Communications Biology*, vol. 8, no. 1, p. 631–632, Apr. 2025, doi: 10.1038/s42003-025-08078-9.

[125] R. L. Carhart-Harris *et al.*, "Neural Correlates of the LSD Experience Revealed by Multimodal Neuroimaging," *Proceedings of the National Academy of Sciences*, vol. 113, no. 17, pp. 4853–4858, Apr. 2016, doi: 10.1073/pnas.1518377113.

[126] R. B. Kargbo, "Unveiling Reality: Psychedelics, Neural Filtering, and the Future of Psychiatric Medicine," *ACS Medicinal Chemistry Letters*, vol. 16, no. 4, pp. 500–503, Apr. 2025, doi: 10.1021/acsmedchemlett.5c00103.

[127] P. R. Corlett, G. Horga, P. C. Fletcher, B. Alderson-Day, K. Schmack, and A. R. Powers, "Hallucinations and Strong Priors," *Trends in Cognitive Sciences*, vol. 23, no. 2, pp. 114–127, Feb. 2019, doi: 10.1016/j.tics.2018.12.001.

[128] L. Zhang *et al.*, "Low-Frequency rTMS Modulates Small-World Network Properties in an AVH-related Brain Network in Schizophrenia," *Frontiers in Psychiatry*, vol. 16, p. 1578072–1578073, Apr. 2025, doi: 10.3389/fpsyt.2025.1578072.

[129] M. S. E. Sendi *et al.*, "Abnormal Dynamic Functional Network Connectivity Estimated from Default Mode Network Predicts Symptom Severity in Major Depressive Disorder," *Brain Connect.*, vol. 11, no. 10, pp. 838–849, 2021, doi: 10.1089/BRAIN.2020.0748.

[130] A. A. T. Simone Reinders, A. T. M. Willemsen, H. P. J. Vos, J. A. Den Boer, and E. R. S. Nijenhuis, "Fact or Factitious? A Psychobiological Study of Authentic and Simulated Dissociative Identity States," *PLOS One*, vol. 7, no. 6, p. e39279, Jun. 2012, doi: 10.1371/journal.pone.0039279.

[131] E. M. Vissia *et al.*, "Dissociative Identity State-Dependent Working Memory in Dissociative Identity Disorder: A Controlled Functional Magnetic Resonance Imaging Study," *Bjpsych Open*, vol. 8, no. 3, p. e82, May 2022, doi: 10.1192/bjo.2022.22.

[132] H. Merckelbach, G. J. Devilly, and E. Rassin, "Alters in Dissociative Identity Disorder," *Clinical Psychology Review*, vol. 22, no. 4, pp. 481–497, May 2002, doi: 10.1016/S0272-7358(01)00115-5.

[133] A. A. T. S. Reinders, "Cross-Examining Dissociative Identity Disorder: Neuroimaging and Etiology on Trial," *Neurocase*, vol. 14, no. 1, pp. 44–53, Feb. 2008, doi: 10.1080/13554790801992768.

[134] A. A. T. S. Reinders *et al.*, "Aiding the Diagnosis of Dissociative Identity Disorder: Pattern Recognition Study of Brain Biomarkers," *British Journal of Psychiatry*, vol. 215, no. 3, pp. 536–544, Sep. 2019, doi: 10.1192/bjp.2018.255.

[135] E. Vermetten, C. Schmahl, S. Lindner, R. J. Loewenstein, and J. D. Bremner, "Hippocampal and Amygdalar Volumes in Dissociative Identity Disorder," *American Journal of Psychiatry*, vol. 163, no. 4, pp. 630–636, Apr. 2006, doi: 10.1176/ajp.2006.163.4.630.

[136] S. Chalavi *et al.*, "Abnormal Hippocampal Morphology in Dissociative Identity Disorder and Post-traumatic Stress Disorder Correlates with Childhood Trauma and Dissociative Symptoms," *Human Brain Mapping*, vol. 36, no. 5, pp. 1692–1704, Dec. 2014, doi: 10.1002/hbm.22730.

[137] Y. R. Schlumpf *et al.*, "Dissociative Part-Dependent Biopsychosocial Reactions to Backward Masked Angry and Neutral Faces: An fMRI Study of Dissociative Identity Disorder," *Neuroimage: Clinical*, vol. 3, pp. 54–64, 2013, doi: 10.1016/j.nicl.2013.07.002.

[138] M. N. Modesti, L. Rapisarda, G. Capriotti, and A. Del Casale, "Functional Neuroimaging in Dissociative Disorders: A Systematic Review," *Journal of Personalized Medicine*, vol. 12, no. 9, p. 1405–1406, Aug. 2022, doi: 10.3390/jpm12091405.

[139] W. J. Clancey, "The Strange, Familiar, and Forgotten: An Anatomy of Consciousness," *Artificial Intelligence*, vol. 60, no. 2, pp. 313–356, Apr. 1993, doi: 10.1016/0004-3702(93)90007-X.

[140] E. Selinger, "Reality+: Virtual Worlds and the Problems of Philosophy," *The Philosophers' Magazine*, no. 98, pp. 110–113, 2022, doi: 10.5840/tpm20229875.

[141] G. Aston-Jones and J. D. Cohen, "AN INTEGRATIVE THEORY of LOCUS COERULEUS-NOREPINEPHRINE FUNCTION: Adaptive Gain and Optimal Performance," *Annual Review of Neuroscience*, vol. 28, no. 1, pp. 403–450, Jul. 2005, doi: 10.1146/annurev.neuro.28.061604.135709.

[142] A. Cleeremans, "Computational Correlates of Consciousness," *Progress in Brain Research*, vol. 150. pp. 81–98, 2005.

[143] J. K. O'Regan and A. Noë, "A Sensorimotor Account of Vision and Visual Consciousness," *Behavioral and Brain Sciences*, vol. 24, no. 5, pp. 939–973, Oct. 2001, doi: 10.1017/S0140525X01000115.

[144] D. A. Shin and M. C. Chang, "Consciousness Research through Pain," *Health Care*, vol. 13, no. 3, p. 332–333, Feb. 2025, doi: 10.3390/healthcare13030332.

[145] H. W. Harris, "Qualia as Query Act, the Phenomenology of Predictive Error Coding," *Frontiers in Psychology*, vol. 16, p. 1531269–1531270, Apr. 2025, doi: 10.3389/fpsyg.2025.1531269.

[146] S. M. Fleming and N. Shea, "Quality Space Computations for Consciousness," *Trends in Cognitive Sciences*, vol. 28, no. 10, pp. 896–906, Oct. 2024, doi: 10.1016/j.tics.2024.06.007.

[147] C. M. Pennartz, "What Is Neurorepresentationalism? From Neural Activity and Predictive Processing to Multi-Level Representations and Consciousness," *Behavioural Brain Research*, vol. 432, p. 113969–113970, Aug. 2022, doi: 10.1016/j.bbr.2022.113969.

[148] A. Sheth, K. Roy, and M. Gaur, "Neurosymbolic AI – Why, What, and How." Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/2305.00813

[149] B. C. Colelough and W. Regli, "Neuro-Symbolic AI in 2024: A Systematic Review," *Lnsai@ijcai*, 2025, doi: 10.48550/ARXIV.2501.05435.

[150] W. Lotter, G. Kreiman, and D. Cox, "A Neural Network Trained for Prediction Mimics Diverse Features of Biological Neurons and Perception," *Nature Machine Intelligence*, vol. 2, no. 4, pp. 210–219, Apr. 2020, doi: 10.1038/s42256-020-0170-9.

[151] P. J. Blazek and M. M. Lin, "Explainable Neural Networks That Simulate Reasoning," *Nature Computational Science*, vol. 1, no. 9, pp. 607–618, Sep. 2021, doi: 10.1038/s43588-021-00132-w.

[152] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A Survey on Neural Network Interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, Oct. 2021, doi: 10.1109/TETCI.2021.3100641.

[153] A. Jansma, P. A. M. Mediano, and F. E. Rosas, "The Fast Möbius Transform: An Algebraic Approach to Information Decomposition," *Corr*, 2024, doi: 10.48550/ARXIV.2410.06224.

[154] A. I. Luppi *et al.*, "General Anaesthesia Reduces the Uniqueness of Brain Connectivity across Individuals and across Species." Accessed: Jun. 21, 2025. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.11.08.566332

[155] R. Scodellaro, A. Kulkarni, F. Alves, and M. Schröter, "Training Convolutional Neural Networks with the Forward-Forward Algorithm." Accessed: Jun. 21, 2025. [Online]. Available: https://arxiv.org/abs/2312.14924

[156] R. Prakki, "Active Inference for Self-Organizing Multi-LLM Systems: {a} Bayesian Thermodynamic Approach to Adaptation," *Corr*, pp. 331–341, 2024, doi: 10.48550/ARXIV.2412.10425.

[157] PubMed, "Hybrid Predictive Coding: Inferring, Fast and Slow," *Plos Comput. Biol.*, vol. 19, no. 8, p. e1011280, Aug. 2023, doi: 10.1371/journal.pcbi.1011280.