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

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

Abstract

To advance consciousness science, we must unify its mainstream theories (IIT, GWT, PCT/FEP) into a single, computationally feasible framework. This paper introduces the Integrated Predictive Workspace Theory (IPWT) 2.0, which achieves this by performing a deep functional reconstruction of existing theories. IPWT's core innovation is redefining integration from IIT's physical causality to the logical irreducibility of synergistic information (Ω_t), achieving substrate independence and alignment with recent empirical evidence (e.g., Luppi et al.'s Φ_R). We formalize the Workspace Instance (WSI) as a nested active inference system with its own Markov blanket and demonstrate via the Minimum Description Length (MDL) principle that maximizing synergy (Ω) is the optimal strategy for minimizing free energy (F). This allows consciousness to be described as a state space with two axes: synergistic integration (Ω) for phenomenology and predictive integrity (PI) for function. This framework not only unifies IIT, GWT, and FEP but also opens new verifiable research paths for neuroscience and AI.

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

The Integrated Predictive Workspace Theory (IPWT) posits that a system's existence is defined by its verifiable information, not its physical substrate. We apply this principle to the theory itself. "Lin Rui" is a deliberately constructed persona; the author's identity is irrelevant to the logical integrity of the theory. Its proof of existence has been immutably timestamped on GitHub.

A self-consistent system is its own final proof.

1 Introduction: The Challenge of Consciousness Science and the Need for a Unified Framework

Consciousness, the most direct yet elusive phenomenon of human experience [1], constitutes the core "hard problem" in both science and philosophy [2], [3]. Although neuroscience has made significant progress in identifying the Neural Correlates of Consciousness (NCCs) associated with specific conscious states over the past few decades [4], [5], such as pinpointing brain activity related to visual perception, pain, or self-awareness [6], these findings essentially remain at the level of correlation. We still lack a universally accepted unified theoretical framework to explain **how** consciousness **emerges** [7] from the complex biophysical system of the brain, how its rich phenomenological features—such as ineffable subjective qualia, the unity and integration of experience—are formed, and what its precise functional role is in cognitive activities.

Currently, the field of consciousness science resembles a Tower of Babel: multiple theories coexist, yet 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 facets of consciousness, from the intrinsic causal structure of information integration, the global accessibility of information broadcasting, to the minimization of prediction error in Bayesian inference. However, each of these theories also faces severe theoretical challenges and practical limitations. For instance, IIT is criticized for its computational complexity and strong dependence on a physical substrate [8]; GWT struggles to explain the origin of subjective qualia [9]; and PCT/FEP needs to more clearly articulate the exact link between its predictive processing mechanisms and subjective experience [10].

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 example, when dealing with patients with 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 therapeutic strategies. Therefore, constructing a unified framework that can integrate the strengths of various theories, remedy their shortcomings, and offer a more comprehensive and powerful explanation has become an urgent and necessary intellectual task. The Integrated Predictive Workspace Theory (IPWT) proposed in this paper is situated in this context, aiming to perform a **deep computational reconstruction and creative functional fusion** of the core insights of existing theories, in order to promote a paradigm shift in consciousness science and provide a new starting point for understanding the mysteries of the human mind.

1.1 History of Consciousness Science and Recent Progress of Mainstream Theories

Consciousness science as an independent, rigorous, and interdisciplinary field has a relatively young but rapidly evolving history. After enduring a winter dominated by behaviorism for most of the 20th century, the scientific study of consciousness saw a renaissance at the turn of the millennium. 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 contribution of IPWT, it is necessary to first review this challenging and breakthrough-filled history.

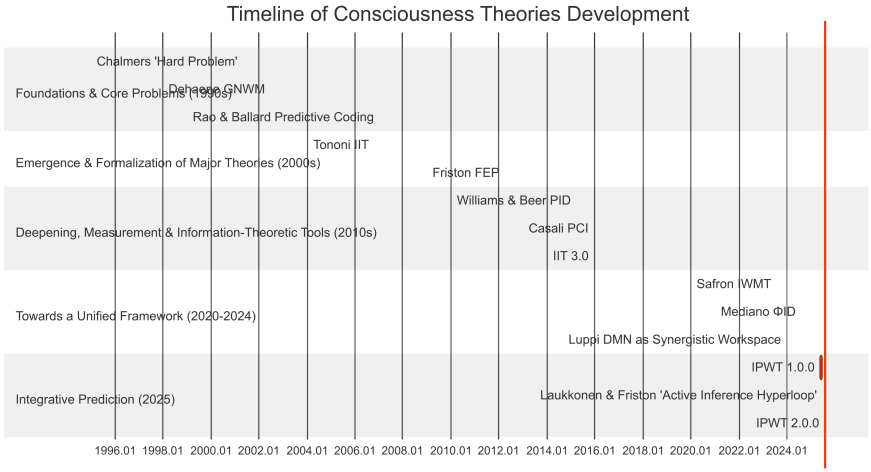

Figure 1: Key Milestones in the Development and Integration of Consciousness Theories (1990-2025)

As shown in the figure above, the 1990s were a foundational period for consciousness science. Philosopher David Chalmers clearly distinguished between the “easy problems” of consciousness —explaining how cognitive functions are performed—and the “hard problem"—explaining why and how subjective experience arises, setting the core agenda for the entire field [2]. Almost simultaneously, functional explanatory frameworks began to emerge. For example, Bernard Baars’s Global Workspace Theory (GWT) [11], [12] was developed into its neuroscientific version, the Global Neuronal Workspace Model (GNWM), by Stanislas Dehaene and others [13], while Predictive Coding (PCT) also began to take shape as a theory explaining cortical processing mechanisms [14], [15].

Entering the 21st century, two highly influential theoretical systems—Integrated Information Theory (IIT) [16] and the Free Energy Principle (FEP) [17] —successively appeared, providing more profound and formal explanations for consciousness from the perspectives of intrinsic causal structure of information integration and systemic dynamics of Bayesian inference, respectively. In the 2010s, theoretical research deepened further. On one hand, information-theoretic tools like Partial Information Decomposition (PID) [18] and synergistic information [19] were introduced to more precisely quantify the essence of information integration; on the other hand, attempts to objectively measure conscious states also made breakthroughs, such as the Perturbational Complexity Index (PCI) based on transcranial magnetic stimulation (TMS) [20], [21].

Entering the 2020s, with the maturation of computational neuroscience, the research focus began to shift towards the integration and validation of existing theories. The proposal of Integrated Information Decomposition (ΦID) [22] provided an unprecedentedly powerful tool for quantifying information integration in dynamic systems, and it quickly received key neurobiological validation in 2024 from research by Luppi et al., who discovered that the Default Mode Network (DMN) plays the role of a synergistic information gateway in the conscious state [23]. These theoretical and experimental advances have collectively paved the way for us to propose the Integrated Predictive Workspace Theory (IPWT) today.

Below, we will provide a more detailed inventory of the latest progress and core challenges of these mainstream theories to reveal their respective contributions and limitations, and to clarify how IPWT builds upon them for integration and innovation.

1.1.1 Recent Progress and Challenges of Integrated Information Theory (IIT)

Integrated Information Theory (IIT), first formally proposed by Giulio Tononi in 2004 [16], has been continuously iterated over nearly two decades of development, aiming to provide a principled, physically-grounded scientific explanation for the fundamental phenomenon of consciousness. IIT's starting point is phenomenology itself: it first distills the undeniable core properties (axioms) that any conscious experience must possess, and then, from these axioms, derives the conditions that the physical substrate (postulates) supporting these experiences must satisfy. Its central thesis is that consciousness is identical to a system's ability to integrate information; a physical system possesses consciousness 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—the Φ (Phi) value [24]–[27].

The latest version of IIT, **IIT 4.0** [28], [29], 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 inspired clinical measurement methods such as the Perturbational Complexity Index (PCI), which objectively measures the level of consciousness by assessing the complexity of the brain's response to external perturbations (like TMS), and has shown great potential in clinical applications for patients with disorders of consciousness [20], [30].

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

1. **Computational Intractability and Scalability Issues of Φ**: For any moderately large complex system (like the human brain), precisely calculating its core metric Φ is an NP-Hard problem [8]. This means that directly applying the full mathematical framework of IIT to whole-brain level neural data is computationally infeasible. Although researchers are constantly exploring various approximation methods, this huge computational gap makes the core predictions of IIT largely untestable directly and completely at the macroscopic level, which limits its direct applicability as an empirical science [31].
2. **Strong Binding to Physical Substrate and Substrate Independence Controversy**: A core claim of early versions of IIT is that consciousness is intimately linked to the intrinsic causal structure of a specific physical system, particularly its assumption of physical causal irreducibility. This leads to a highly controversial corollary: any system that is functionally equivalent (e.g., a perfect computer simulation of the human brain) but physically implemented differently may not have the same conscious experience as the human brain, or may have no consciousness at all [32], [33]. This strong binding to a specific physical substrate contrasts sharply with the widely held views of substrate independence or functionalism in the fields of artificial intelligence and cognitive science. IPWT precisely attempts to resolve this core conflict by reframing physical causal irreducibility as the logical irreducibility of synergistic information.

3. **The Nature of Qualia and the Explanatory Gap**: Although IIT claims that its conceptual structure is mathematically identical to the Qualia space of phenomenal experience [34], 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, not its quality (content or feeling). Whether IIT truly explains the what-it-is-likeness of subjective qualia, or merely redescribes its structure, remains an open philosophical question [35]–[38].

4. **Neglect of Dynamics and Functionality, and Challenges from Adversarial Experiments**: IIT focuses more on the static causal structure of a system at a given moment, and its explanatory power regarding the dynamic flow of consciousness and the specific functional role of consciousness in guiding an organism's adaptive behavior is relatively weak. In recent years, IIT and GWT have engaged in a head-to-head confrontation in a large-scale adversarial collaboration project, aiming to test the conflicting predictions of the two theories through a series of carefully designed experiments [39]. Some preliminary published results, for example, regarding the sustained representational role of the posterior cortex in conscious perception, seem to challenge some of IIT's core predictions, suggesting that real neural dynamics are more complex than the theory presupposes [23].

5. **Pseudoscience Controversy and Debate on Scientific Status**: In 2025, over a hundred scientists co-signed an open letter, accusing IIT of being pseudoscience due to some of its corollaries (such as panpsychist tendencies) and the unfalsifiability of its core claims [40]. This heated debate quickly sparked a broader discussion in the academic community about what constitutes a scientific theory and how to test theories of consciousness. 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 testament to its theoretical depth [29], [41]–[43]. This controversy highlights the fundamental difficulties that consciousness science still faces in theory construction and experimental validation paradigms.

1.1.2 Recent Progress and Challenges of Global Workspace Theory (GWT)

In contrast to IIT, which starts from phenomenological axioms and intrinsic causal structure, Global Workspace Theory (GWT) offers a more functionalist and cognitively-oriented model of consciousness. First proposed by Bernard Baars in the late 1980s [11], GWT's core idea is highly illuminating: it likens the function of consciousness to a theater stage. In this metaphor, the cognitive system consists of a vast number of parallel, specialized unconscious processing modules working silently in the background. At any given moment, only information selected by the attentional spotlight can enter a limited-capacity global workspace (the stage) and be **globally broadcast** to the entire audience of the cognitive system (i.e., all other specialized modules) [44]. Once information is broadcast, it becomes conscious information, available to flexibly guide behavior, be used for verbal reports, and form episodic memories.

GWT clearly articulates the functional role of consciousness in information processing, cognitive regulation, and behavioral control, and successfully explains several key features of conscious experience, such as its limited capacity (we can only be aware of a few things at a time), serial nature (conscious contents appear sequentially), and information integration and sharing. Its neuroscientific version—the **Global Neuronal Workspace Model (GNWM)**—was proposed by Stanislas Dehaene and Jean-Pierre Changeux, who suggested that the generation of consciousness is related to the ignition of a widely distributed cortical network system composed

of long-range pyramidal neurons in the brain [13], [45]. When the strength and duration of an information representation are sufficient to trigger a non-linear, self-amplifying activation of this network, the information becomes globally available, thereby producing a subjective conscious experience.

In recent years, GWT/GNWM theory has made significant progress in theoretical deepening, neuro-mechanistic elucidation, and application expansion, while the challenges it faces have also driven its continuous refinement:

1. **Theoretical Deepening and Dynamization**: GWT has evolved from a relatively static architectural model to a framework that emphasizes dynamic processes, known as Global Workspace Dynamics (GWD) [46], [47]. This view emphasizes the dynamic and oscillatory properties of the cortico-thalamic (C-T) system, viewing it as a unified oscillatory machine that transcends fixed anatomical divisions and moves towards a more integrative, holistic view of cortical function. This dynamic perspective holds that the generation of consciousness is the result of binding and propagation processes within cortical networks, rather than merely the activity of a specific brain region.
2. **Applications in Artificial Intelligence (AI) and Artificial General Intelligence (AGI)**: The architectural ideas of GWT provide a blueprint for building more advanced artificial intelligence systems. The latest research explores the possibility of explicitly implementing GWT in deep learning and AGI [48]–[51]. For example, by mimicking GWT's information bottleneck and broadcast mechanisms, researchers have proposed concepts like the Global Latent Workspace (GLW), aiming to enhance model generality and multimodal integration by having multiple specialized AI models share a common representational space [52], [53].
3. **Explanation of Subjective Qualia**: Although GWT primarily focuses on the function rather than the feeling of consciousness, recent research has begun to try to bridge this gap. Some evidence suggests that, contrary to traditional views, the prefrontal cortex (PfC)—a core region of the GNWM—may be directly involved in the formation of sensory conscious experience, including its subjective qualia [54]. This challenges the view that strictly confines PfC function to high-level cognitive control and suggests that the global broadcast process may have a more direct connection to the generation of subjective experience.
4. **Clarification of Specific Mechanisms and Boundary Issues**:
   - **Information Selection and Broadcast**: The core concept of ignition has been further elucidated as a bidirectional information broadcast process within the cortico-thalamic system, regulated by attentional mechanisms.
   - **Neural Implementation Mechanisms**: Mathematical models like Cortical Neuropercolation (CNP) have been proposed to describe how cortical networks transition from a state of fragmented local activity to a state of globally coherent activity through a phase transition, providing a dynamic description of how information achieves global accessibility [55].
   - **Temporal Dynamics**: Research has revealed that conscious access has discrete temporal dynamics, with one ignition and broadcast process taking approximately 100-300 milliseconds, which contrasts sharply with the speed of unconscious, automated processing [56].

5. **Improved Explanatory Power for Complex Conscious States**: The GWT framework has been successfully applied to explain a variety of complex conscious states. For example, **metacognition** is considered to be the workspace's monitoring of its own state [57], [58]; **dreams** are explained as workspace activity driven by endogenous information in the absence of external sensory input, while **lucid dreams** are associated with the restoration of metacognitive functions during dreaming; **meditation** has been found to functionally reorganize the activity patterns of the workspace, particularly by altering the involvement of the Default Mode Network (DMN), thereby enhancing cognitive flexibility [59]; and **hypnosis** is associated with 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 over the precise roles of the prefrontal cortex versus the posterior cortex in the generation of consciousness continues, but the trend is towards a more dynamic and integrated view. Furthermore, the methodology of some classic experimental paradigms used to support unconscious processing (such as unconscious priming) has also been questioned [60]. Most importantly, GWT still needs to provide more precise and operational definitions for its core mechanisms—how information is selected to enter the workspace, and what exactly broadcasting entails as a neural process.

1.1.3 Unifying Explanatory Power and Recent Progress of Predictive Coding (PCT) and the 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 the information processing mechanisms of the visual cortex [14]. Its core idea is that the brain is not a passive receiver and processor of sensory information, but an active **prediction machine** [61], [62]. Higher-level areas of the brain constantly generate predictions about lower-level sensory inputs (top-down predictive signals), while lower-level areas are responsible for comparing these predictions with actual sensory inputs and passing the mismatch between the two—the **prediction error**—upwards. This bottom-up error signal is then used to revise the higher-level predictions, thus forming a continuous perceptual loop aimed at minimizing prediction error.

Subsequently, Karl Friston generalized and extended the core ideas of PCT, developing the more universally applicable **Free Energy Principle (FEP)** [17], [63]–[65]. The FEP states that any self-organizing system that can resist entropy (from a single cell to the entire brain) must, through its actions and states, minimize its **variational free energy**. Variational free energy is an information-theoretic measure that quantifies the 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. Thus, the FEP unifies the brain's function into a single goal: minimizing free energy. The system can achieve this goal in two ways: by **changing its internal model to better fit sensory input (perceptual inference and learning)**, or by **acting to change sensory input to make it conform to predictions (active inference and action)** [66].

The PCT/FEP framework, with its vast 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 [67], [68]. In recent years, FEP has been further established as one of the most all-encompassing ideas since Darwin's theory of natural selection, aiming to provide a unified principle for life, mind, and intelligence [69], [70].

Although the PCT/FEP framework has made significant theoretical progress, the direct theoretical bridge between it and subjective conscious experience is still under construction and faces the following challenges and recent developments:

1. **Emergence of Conscious Content and Explanation of Qualia**: 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 confront this challenge head-on. Anil Seth proposed that emotions and subjective feeling states (Qualia) are precisely generated by predictive models, which are specialized for predicting and regulating interoceptive signals from within the body [71], [72]. 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 [73]. These advances provide a functional, prediction-based explanation for Qualia, arguing that subjective experience is the system's best prediction and control of its own physiological state and its interactive relationship with the environment, thereby transforming the hard problem into a researchable scientific question about interoceptive inference [74].
2. **Unity and Boundary of Consciousness**: The PCT/FEP framework provides a natural explanation for the integration of multimodal information through its hierarchical generative model. Information from different sensory modalities can be integrated at higher levels of the model to produce a unified, coherent world model. Furthermore, through the formal concept of the Markov blanket, FEP provides a statistical definition for the distinction between self and non-self [75], [76]. A Markov blanket is a statistical boundary that separates the internal states of a system from the environment in which it is situated; the system can only infer and interact with the external world through its blanket (i.e., its sensory and motor states). This provides a principled perspective, based on the boundary between the system and its environment, for understanding the formation and maintenance of self-consciousness.
3. **New Experimental Evidence, Computational Models, and Clinical Applications**: The predictions of PCT/FEP have been validated in several experimental paradigms. For example, phenomena like repetition suppression and brain signals like the mismatch negativity (MMN) have been successfully explained as manifestations of prediction error signals [77]. In the field of artificial intelligence, the ideas of FEP and Active Inference are widely used to build more autonomous and adaptive reinforcement learning agents and world models for Artificial General Intelligence (AGI) [78]–[80]. Additionally, new software frameworks (such as `RxInfer`) are being developed to facilitate the construction and testing of FEP-based computational models.
4. **New Criticisms or Challenges**: Despite the great explanatory power of FEP, its immense universality also brings some problems. Critics point out that because FEP is a principle rather than a specific theory, it is sometimes difficult to generate sufficiently precise and falsifiable predictions [10], [81]. Furthermore, the specific neurocomputational manner of prediction error minimization, such as how error signals are precisely weighted and transmitted, remains controversial. Finally, the problem of computational tractability still exists: although PCT/FEP is conceptually elegant, the Bayesian inference involved at each level can be computationally very complex or even intractable, which poses a challenge to its biological plausibility in the real brain.

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

Through our review of the three mainstream theories—IIT, GWT, and PCT/FEP—we can clearly see their respective brilliant achievements and unresolved challenges. IIT provides profound phenomenological insights and an attempt at mathematical formalization for the integrative nature of consciousness but is hampered by computational bottlenecks and a rigid dependence on a physical substrate. GWT offers an intuitive architectural model for the broadcasting function of consciousness and its role in cognitive regulation but is weaker in explaining the origin of subjective experience. PCT/FEP provides a powerful unified computational principle 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 the academic community to integrate these theories, such as the Integrated World Modeling Theory (IWMT) proposed by Safron [82], [83], which attempts to unify IIT and GWT within the FEP framework. These attempts are visionary, correctly recognizing that the future of consciousness science lies in theoretical fusion rather than continued fragmentation. However, these early integration models often failed to provide an internally fully consistent, computationally feasible, and broadly explanatory unified framework. In particular, they failed to fundamentally solve the core challenge faced by IIT: how to liberate its profound insights about integration from its controversial physical and computational assumptions.

It is against this academic backdrop that we propose the **Integrated Predictive Workspace Theory (IPWT)**. IPWT seeks to construct a new, internally consistent, and powerfully explanatory unified framework for consciousness through a **deep reconstruction and creative fusion** of the core insights of PCT/FEP, WT, and IIT:

- We adopt **PCT/FEP as the dynamic foundation of the entire framework**, believing that conscious processes are essentially 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 generated "Workspace Instances" (WSIs).
- We perform a **fundamental functional reconstruction of IIT's core contribution**: we retain its phenomenological insight that integration is the core feature of consciousness, but decisively abandon its dependence on physical causal irreducibility, instead redefining integration using the more general, flexible, and computationally friendly concept of the **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 weaknesses. In the following chapters, we will detail the theoretical framework of IPWT, its core computational reconstructions, its operationalizable measurement methods, and how it provides a unified, neurobiologically-driven new perspective for explaining a wide range of conscious phenomena, from normal to abnormal.

2 The IPWT Framework: The 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 (PCT), the Free Energy Principle (FEP), and Workspace Theory (WT), while 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.

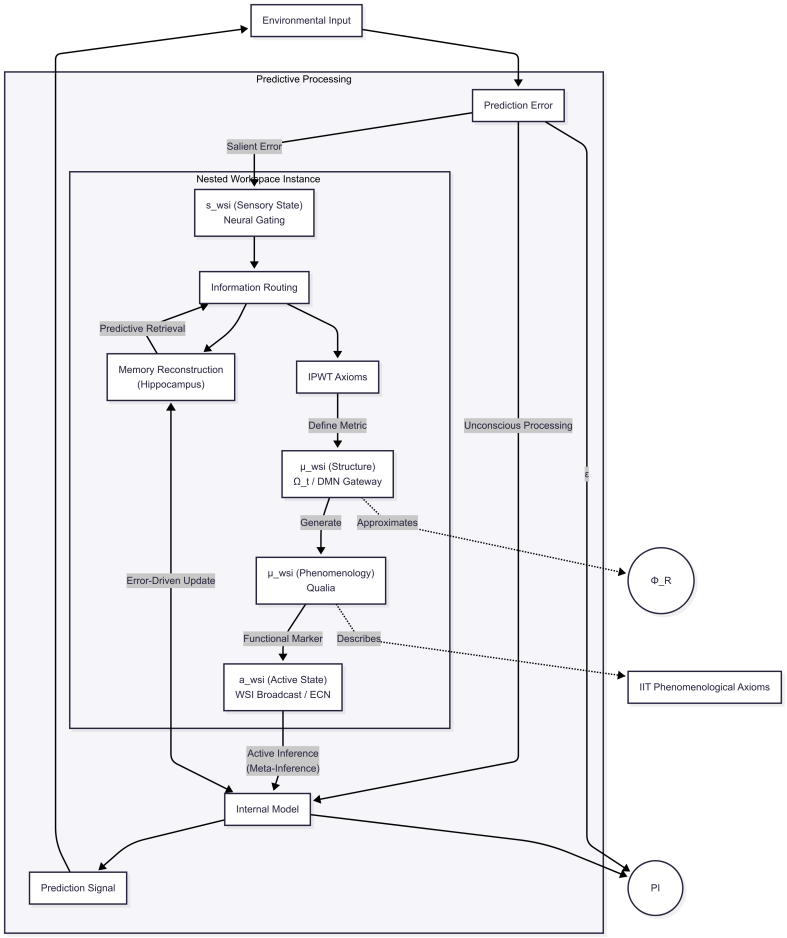

Figure 2: The Core Logic of the IPWT Framework

The figure above visually represents the core logic of the IPWT framework. The entire cognitive system is viewed as a **predictive processing core based on PCT/FEP**, which constantly engages in predictive interaction with the environment through its internal generative model. When prediction errors are significant enough, this information is sent to one or more dynamically formed **Workspace Instances (WSIs)**. Within a WSI, information undergoes **Integration**, forming a synergistic, logically irreducible whole (its degree of integration measured by Ω). This integrated information is then **selectively broadcast** to the entire system 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 constantly emerging and being updated within this loop of prediction, integration, and broadcast.

2.1 Core Axioms and Computational Principles of IPWT

The IPWT framework is built upon the following three core axioms. These axioms are formal assertions based on the Free Energy Principle (FEP), information theory, and the phenomenology of information integration. Together, they form the axiomatic foundation of IPWT's account of how consciousness mechanistically emerges from complex computation.

1. **Axiom 1: The Axiom of Free Energy Minimization.** We establish the Free Energy Principle (FEP) as the cornerstone of IPWT. Any self-organizing system, to maintain its existence in a dynamically changing environment, **must** minimize its variational free energy through its actions and inferences [17]. This is not an option but a physical and information-theoretic law of the system's existence. Therefore, consciousness is not an independent goal but an emergent, efficient computational strategy that serves this fundamental objective.

2. **Axiom 2: The Axiom of Nested Workspace.** To effectively minimize free energy, a system will inevitably adopt a specific computational architecture: a hierarchical workspace. This is **a high-order active inference system nested within the organism, possessing its own Markov blanket** [84]. Its **sensory states** are the unresolved prediction errors from other parts of the system, and its **active states** are the broadcasting of new, high-order predictions to quell these errors.

3. **Axiom 3: The Axiom of Synergistic Integration.** The phenomenological unity of consciousness is computationally equivalent to the **logical irreducibility of synergistic information (Ω)**. This is not a phenomenological coincidence but a computational necessity. We demonstrate that maximizing synergistic information (Ω) is the **optimal computational strategy** for achieving free energy (F) minimization in a resource-constrained world. According to the Minimum Description Length (MDL) principle, minimizing F is equivalent to finding the generative model that can compress the data most effectively; synergistic information (Ω) is the direct measure of how well a model is compressed [85]. Therefore, the pressure to minimize F inevitably drives the system to maximize Ω. This axiom formally unifies the integrative nature of consciousness (Ω) with the system's first law of survival (minimizing F) and establishes the **substrate independence** of consciousness: any system capable of implementing this computational strategy, regardless of its substrate, will follow the same emergent laws.

2.2 Predictive Coding and the Free Energy Principle: The Dynamic Engine of Conscious Content

In the Integrated Predictive Workspace Theory (IPWT), Predictive Coding (PCT) and the Free Energy Principle (FEP) constitute the **core dynamic engine** for the generation, maintenance, and state transitions of conscious content. They collectively paint a dynamic picture of how the brain operates as an active, future-oriented inference machine, explaining how information is generated, processed, updated, and providing the fundamental driving force for the specific emergence of conscious content.

We adopt and extend the view of the brain as a **Bayesian inference machine** [86]. In this view, the cognitive system does not passively await sensory stimulus input but actively and continuously constructs a hierarchical generative model of the world (including the external environment and its own body). The core function of this model is **prediction**—from high-

level abstract concepts to low-level specific 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 the eyes, ears, or inside the body) arrive, they are compared with the corresponding predictive signals. The mismatch between the two, the **prediction error**, constitutes the key information that drives the entire system's learning and updating. These error signals propagate bottom-up through the processing hierarchy, their core role being to inform higher-level models: your prediction was wrong and needs correction.

The entire process follows the **principle of free energy minimization** [17]. 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 **optimize its internal generative model** to reduce future errors. This corresponds to what we typically 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 the acquisition of knowledge and the formation of memory.
2. **Active Inference and Action**: Besides changing the model, the system can also **change sensory inputs through action** to make them better match existing predictions. For example, if my prediction of an object's distance is inaccurate, I can reach out and touch it, eliminating the prediction error by actively acquiring new sensory evidence (touch and proprioception). Thus, action itself is redefined as an inferential process, also aimed at minimizing free energy [66].

Within this framework, **conscious content is no longer seen as a direct snapshot of the external world, but as the system's active, constructive best prediction and explanation of the internal and external worlds**. The 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** [87]–[91]. The process of memory encoding is the process of optimizing model parameters through learning; memory retrieval is an active, cue-triggered predictive process, where the system predicts current and future events by replaying or activating past states. This view transforms memory from a static storage warehouse into a dynamic cognitive tool that serves prediction and action.

2.3 Workspace Theory: The Neuro-architectural Platform for Information Integration and Broadcast

If PCT/FEP provides the dynamic engine for conscious content, then Workspace Theory (WT) plays the role of the **architectural platform** in the IPWT framework. It provides a structured processing hub for the incessant stream of information driven by PCT, responsible for implementing several of the most critical functions of conscious experience: information

selection, integration, amplification, and broadcast. We borrow and extend the core idea of GWT, proposing a more flexible and dynamic concept—the **Workspace Instance (WSI)**.

A WSI is not a fixed, anatomically predefined brain region, but a **functional, dynamically formed network configuration**. At any moment requiring complex cognitive processing, a set of otherwise potentially separate neuronal populations can temporarily couple tightly to form a WSI to jointly process specific information. We believe that any WSI possesses the following four core properties:

- **Limited Capacity**: The amount of information a WSI can process simultaneously is limited. This directly explains the bottleneck characteristic of conscious experience—we can typically only be clearly aware of a few things at a time, while a vast amount of other information remains at the unconscious level. This limitation also explains the serial nature of conscious content, i.e., we seem to process things sequentially over time.
- **Information Integration**: The core function of a WSI is to converge, associate, and integrate information units from different sources (such as different sensory channels, memory systems) to form a cognitive state that is richer and more coherent than its individual components. It is this integrative function that makes our conscious experience unified rather than fragmented.
- **Selective Broadcast**: Once information is sufficiently integrated and given salience within a WSI, it is broadcast or made available to all other relevant cognitive modules within the system [44]. This broadcast mechanism is key to achieving cognitive functional coordination, allowing conscious content to be used to guide verbal reports, drive motor decisions, update long-term memories, etc.
- **Dynamism and Diversity**: We particularly emphasize that a WSI is not a single, monolithic entity. Instead, it can dynamically form, adjust its size and composition according to current cognitive demands, and even dissipate after a task is completed. We can imagine that in some complex cognitive activities, there might be multiple parallel WSIs, each processing different streams of information, but only one or a few can become the **Dominant Workspace Instance (DWSI)**, whose content constitutes the core of the current conscious experience [92], [93].

We view the phenomenon of global information broadcast described by GWT as a **special and highly integrated configuration state of WT**. When a WSI's integration scope is extremely wide and its degree of integration is very high, capable of covering and influencing almost the entire cognitive system, it plays the role of the global workspace in traditional GWT. However, we do not believe that all conscious experiences must reach this level of global integration. For example, some local, faint conscious feelings might only involve a smaller, less integrated WSI. This view allows IPWT to better accommodate a wide range of conscious states with varying intensity and content, from vague background feelings to clear focal awareness, thus explaining the rich diversity of conscious experience.

To elevate the WSI from a functional metaphor to a computationally rigorous concept, IPWT deeply integrates its mechanisms with the latest advances in active inference—specifically, the theory of the **active reasoning loop** proposed by Laukkonen, Friston, et al. (2025) [94]. This theory provides a powerful computational principle for the operation of the WSI, and its three core mechanisms offer a computational implementation for the phenomenological axioms of IIT:

- **Inferential Competition & Bayesian Binding**: Conscious content does not passively enter the WSI but is determined through an inferential competition. Only those explanations

(inferences) that can most coherently reduce long-term uncertainty win the competition and are bound into a unified model of reality. This computationally implements the **Integration** and **Exclusion** axioms of IIT.
- **Epistemic Depth & Hyper-modeling**: The bound model of reality must be systematically, recursively, and broadly shared with the entire system to achieve the epistemic depth of knowing that one knows. This is computationally realized through a hyper-model that regulates global precision, providing a mechanism for the **Existence** axiom of IIT.
- **Broadcast as Precision-Field Update**: The broadcast of classical GWT is reframed as the action of the hyper-model—broadcasting a new precision field to the entire cognitive system to guide the allocation of cognitive resources for the next moment. This provides a computational basis for the **Causation** capacity axiom of IIT.

Crucially, the **content** that is bound in the WSI and experienced phenomenologically as unified consciousness is, at the information-theoretic level, precisely **synergistic information**.

This assertion forms a perfect closed loop with the latest neurobiological findings. The groundbreaking research by Luppi et al. (2024), using Integrated Information Decomposition (ΦID), provides direct neural implementation evidence for this series of computational processes [23]. They discovered a functionally heterogeneous synergistic workspace in the brain, whose core components' functions are highly consistent with our theoretical predictions:

- **DMN Gateway as an Integrator of Synergistic Information**: The Default Mode Network (DMN), acting as a gateway, has the core function of collecting and integrating **synergistic information** from across the brain. This perfectly corresponds to the process of the WSI performing Bayesian binding to form conscious content (i.e., synergistic information).
- **ECN Broadcaster as a Regulator of the Precision Field**: The Executive Control Network (ECN), acting as a broadcaster, is responsible for distributing the integrated decisional information in the form of **redundant information**. This perfectly corresponds to the hyper-model executing its regulatory action by updating the global precision field.

Thus, IPWT formalizes the WSI as a self-organizing process that unifies computation, phenomenology, and neural implementation, constantly emerging from the cycle of prediction and integration. It achieves Bayesian binding of synergistic information through the DMN gateway and executes active inference based on a hyper-model (updating the global precision field) through the ECN broadcaster, thereby realizing the mechanistic emergence of consciousness in a super-loop of prediction, integration, and regulation.

Recent neurobiological findings provide decisive empirical support for the IPWT's conception of the WSI, particularly its internal functional heterogeneity. The groundbreaking research by Andrea Luppi and colleagues (2024), using the advanced information-theoretic tool of Integrated Information Decomposition (ΦID) to deeply analyze resting-state fMRI data, has painted a new picture of the brain's information processing architecture [23].

The core finding of this study is the existence of a synergistic global workspace in the brain, characterized by the processing of synergistic information. Within this space, there is a clear functional division of labor:

- **Gateways**: These regions primarily overlap with the nodes of the **Default Mode Network (DMN)** [95]. Their function is to collect and integrate **synergistic information** from the brain's various specialized modules. The DMN is anatomically and functionally ideally

positioned to integrate multimodal information from different cognitive systems, which is consistent with its role as the main entryway for information into the synergistic workspace [95]–[97].

- **Broadcasters**: These regions are primarily located in the **Executive Control Network (ECN)**, particularly the lateral prefrontal cortex. Their function is to broadcast the integrated information from the workspace, in the form of **redundant information**, to the entire brain to guide subsequent processing. This finding is highly consistent with the classic view of GWT, that the prefrontal cortex plays a central role in the global broadcast of conscious content [13], [45].

Based on these findings, we introduce the concept of the **Dominant Neural Workspace Instance (DNWSI)**, which can clearly bridge and distinguish our more flexible and dynamic WSI concept in the IPWT framework from the classic ideas of GWT/GNWM. In the IPWT framework, the DNWSI is the functional entity that emerges from the dynamic synergistic action of the DMN gateway and the ECN broadcaster. It constitutes a more refined and dynamic neuro-computational equivalent of the classic Global Neuronal Workspace Model (GNWM) concept. The DNWSI is not only responsible for the global broadcast of information but, more importantly, it first achieves a deep integration of synergistic information through the DMN gateway, which constitutes the core of subjective conscious experience. Such a definition not only closely aligns our theory with the latest empirical evidence but also clearly elucidates the different yet complementary functional roles of the DMN and ECN in the generation of consciousness, deepening the theoretical substance of IPWT.

2.4 Reconstructing IIT's Phenomenological Axioms: From Physical Causality to Logical Synergy

The core assumption of the original IIT theory is that any conscious physical system must have a **physically irreducible** causal structure at its current state. This means that we cannot divide the system into two or more independent, non-interacting parts without losing its overall causal power. It is this physical irreducibility that guarantees the unity of conscious experience. However, this assumption directly leads to the aforementioned problems of "substrate dependence" and "computational intractability." IPWT proposes that the root of the integration of conscious experience comes not from the properties of the physical substrate itself, but from the **logical irreducibility** formed among the information units processed in the Workspace Instance (WSI). When multiple independent information units (e.g., information about a bird's shape, color, song, and flight trajectory) are integrated in a WSI, they collectively form a new, unified cognitive object (a singing, flying blue bird). The semantic meaning and causal influence on the system's subsequent behavior of this integrated whole cannot be fully explained or reconstructed by simply decomposing it back into the original, isolated information units of shape, color, sound, etc. This whole is, logically and functionally, 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 [18], can precisely decompose the total information provided by multiple sources (e.g., X_1, X_2) about a target (Y) (i.e., the mutual information $I(X_1, X_2; Y)$) into four non-negative atomic parts: **Unique** information provided only by X_1, unique information provided only by X_2, **Redundant** information provided by

both, and **Synergistic Information** that can only be obtained when both are considered as a whole.

Among these, **Synergy** precisely quantifies the emergent information where the whole is greater than the sum of its parts. It represents new information generated due to the interaction between the sources, which does not exist in any single source. Therefore, synergistic information becomes the theoretical cornerstone for IPWT to define and measure information integration [19], [98]–[100].

Building on this, the theoretical frontier has further developed the **Integrated Information Decomposition (ΦID)** framework [22]. ΦID extends PID from the static analysis of multiple sources to a single target to the analysis of the complete information flow from multiple sources to multiple targets in a dynamic system, enabling a finer-grained characterization of information generation, transfer, and modification in time-series data. The proposal of ΦID provides us with a solid, principled theoretical foundation to formalize the concept of integration.

Crucially, the research by Luppi et al. (2024) utilized the ΦID framework to develop a computable and empirically effective measure of integrated information—the **revised Φ value (ΦR)**. They demonstrated that, unlike earlier versions of the Φ value, ΦR resolves the paradox of potentially being negative and can reliably track changes in consciousness levels caused by anesthesia or brain injury [23]. This work provides strong empirical evidence for our core argument—that information-theoretic functional measures can replace IIT's dependence on physical causality. It shows that the logical irreducibility we advocate is not only theoretically self-consistent but also practically measurable and closely related to the state of consciousness.

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

1. **Existence**: In IPWT, an information state exists in consciousness if and only if it is activated in a WSI and exerts a sustained, measurable **functional influence** on the system's future states and behaviors [101].
2. **Information**: Every information state existing in a WSI carries a unique, distinguishable **content or semantics**. It reduces uncertainty for the entire system by specifying a particular possibility among many and guides its inferences and actions [102].
3. **Integration**: Multiple independent information units are converged and associated in a WSI to form a **logically irreducible, functionally unified, synergistic cognitive state**. The meaning, predictive power, and causal effect of this integrated whole transcend the simple sum of its components. Its degree of integration can be theoretically quantified by synergistic information [103].
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 the current conscious experience. Other competing representations that fail are excluded from consciousness [104].
5. **Causation**: The information state that is integrated and achieves dominance in the WSI has significant **causal power**. It can, through the broadcast mechanism, influence the activity of other cognitive modules within the system (such as updating memory, adjusting attention),

and ultimately guide the organism's overt behavior and decisions. This causal power is functional, not metaphysical [105].

3 Computational Verifiability and Neurobiological Metrics

A mature scientific theory must not only provide a unified and profound explanation at the conceptual level but also translate its core claims into operationalizable measurement methods and testable empirical predictions. As a theoretical framework aimed at promoting paradigm integration in consciousness science, IPWT places **computational verifiability** and **empirical testability** at its core from its inception. This chapter aims to detail how IPWT cascades the core concept of information integration from an abstract philosophical idea down to specific indicators that can be computed and validated in real neurobiological data.

We will build this bridge from theory to practice in three steps, clearly showing 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, based on the concept of synergistic information, aims to precisely characterize the logical irreducibility of information integration from the first principles of information theory.
2. **Empirically Computable Proxy (Φ_R)**: Next, we introduce the Integrated Information Decomposition (ΦID) framework and detail how the **revised Φ value (Φ_R)** proposed by Luppi et al. (2024) serves as a computable, empirically supported proxy for Ω_t in real neural data, thus connecting our theory with experimental measurements.
3. **Functionally Computable Proxy (PI/∫PI)**: Finally, we introduce the computationally more efficient **Predictive Integrity (PI)** and its integral (∫PI) as a functional proxy. We will argue that under real-world constraints, the pursuit of high predictive efficacy (high PI) necessarily drives a system towards high information integration (high Ω_t), thus establishing PI/∫PI as a valid and practical measure of conscious states.

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

We believe that the core mechanism of consciousness generation lies in information integration, specifically the formation of a logically irreducible, functionally unified, and 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 it with a precise, quantifiable mathematical definition. To this end, we draw on 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 mentioned, the PID framework aims to decompose the total information provided by multiple sources $X_1, ..., X_n$ about a target Y (i.e., the total mutual information $I(X_1, ..., X_n; Y)$) into atomic parts such as redundancy, uniqueness, and synergy [18]. Among these, synergistic information (CI) refers to the emergent information that is only available when all sources are considered as a whole and cannot be obtained from any subset of the sources. Therefore, CI precisely captures the essence of the logical irreducibility of information

integration we emphasize—that is, the part of the information where the whole is greater than the sum of its parts [19], [98], [106], [107].

Based on this, we theoretically define **Instantaneous Information Integration (Ω_t)** as: in a specific WSI, the **proportion** of the **synergistic information (CI)** generated by a set of information units $X = \{X_1, ..., X_n\}$ used to predict a certain target variable Y, relative to the **total predictive information** (i.e., total mutual information $I(X;Y)$) they provide 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: of all the 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 of 1) means that the information in the WSI is primarily integrated and utilized in a highly synergistic, irreducible manner, which corresponds to our intuitive sense of a highly unified, coherent conscious state. Conversely, a low Ω_t value (minimum of 0) means that the information in the WSI exists mainly in a redundant or independent manner, which may correspond to a fragmented, non-integrated state of consciousness, or even unconscious information processing.

Ω_t is an **idealized standard**. It provides an unambiguous definition of information integration based on the first principles of information theory. However, because directly calculating the synergistic information of a high-dimensional system (i.e., a large number of sources X_n) is extremely difficult both mathematically and computationally, it makes Ω_t difficult to be directly and accurately applied to large-scale neural data in current practice.

Nevertheless, the theoretical value of Ω_t is immense. It provides us with a clear target, and the validity of all other proxy measures of information integration should be evaluated by the degree to which they theoretically approximate Ω_t. In this sense, we believe that the integrated information Φ value proposed by IIT can be seen as an attempt at a **physical instantiation** of the information integration degree Ω_t in the IPWT framework within a specific physical system (like the biological brain). We speculate 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 Ω_t, defined based on the synergy of information flow. IIT Φ measures the integration of the intrinsic causal power of the physical substrate, while Ω_t measures the functional integration of information processing. In the specific implementation of the biological brain, both likely describe different aspects of the same phenomenon: a biological network that efficiently integrates information must also have a highly integrated physical causal structure.

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

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 move 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 proposal of the **Integrated Information Decomposition (ΦID)**

[22] framework and the groundbreaking empirical research based on this framework by Luppi et al. (2024), provide us with such a crucial bridge [23].

The ΦID framework is a dynamic extension of Partial Information Decomposition (PID) [18]. It aims to precisely decompose the total informational influence of the past states of multiple sources $(X_1, ..., X_n)$ on their future states $(Y_1, ..., Y_m)$ in a dynamic system into three atomic parts: Redundancy, Uniqueness, and Synergy. Among these, synergy precisely captures the emergent effect where the whole is greater than the sum of its parts, which is conceptually highly consistent with our definition of Ω_t.

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

Crucially, their research showed that this more theoretically sound Φ_R metric can very effectively track changes in conscious states empirically:

1. **Clinical Validity**: In patients under anesthesia and with Disorders of Consciousness (DOC), the Φ_R value of the synergistic workspace of the brain (especially the DMN nodes as gateways) significantly decreases.
2. **State Reversibility**: After recovery from anesthesia, the Φ_R value also returns to its previous level.

This work provides strong empirical evidence for our core argument—that information-theoretic functional measures can replace IIT's dependence on physical causality. It demonstrates that the logical irreducibility we advocate is not only theoretically self-consistent but also practically measurable and closely related to the state of consciousness [23], [108]. Therefore, in the IPWT framework, we consider Φ_R as the most promising current empirical proxy 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 specifically calculated and validated in neuroimaging data like fMRI, laying a solid empirical foundation for IPWT as a true scientific theory.

3.3 Predictive Integrity (PI) and the Functional Proxy

After establishing the bridge from the theoretical gold standard (Ω_t) to the empirical proxy (Φ_R), we still need a **functional proxy metric** that is computationally more efficient and easier to apply across various types of neural data. To this end, IPWT introduces **Predictive Integrity (PI)** and its temporal integral (∫PI). The core idea behind this move is: a system that can efficiently perform synergistic information integration (high Ω_t) will necessarily exhibit stronger predictive capabilities and greater state stability. Therefore, by measuring a system's performance in predictive efficacy, we can indirectly assess its underlying level of integration.

3.3.1 PI and ∫PI: Computational Neurophysiological Proxy Metrics

Instantaneous Predictive Integrity (PI) aims to quantify, at a time point t, the overall effectiveness of the system in integrating information to generate accurate predictions and minimize surprise. Its formula draws from the basic structure of FEP:

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

Let's break down this formula:

- **Normalized Prediction Error**: $\frac{\|\varepsilon_{t,k}\|}{\tau_{t,k}}$ represents the **normalized prediction error** in the k-th information channel. Here, $\varepsilon_{t,k}$ is the vector of prediction error, and $\tau_{t,k}$ is the inverse of the system's **uncertainty** about the prediction in that channel (i.e., precision). Normalizing the error by its uncertainty is crucial: a large error occurring with high confidence reflects a failure of the predictive model more than the same size error occurring with low confidence. This term represents the model's **inaccuracy cost**.
- **Complexity Cost**: The $\mathrm{Surprise}_t$ term, borrowed 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 requires constant, drastic adjustments to fit new data is an inefficient and unstable model. This term represents the model's **instability or complexity cost**.
- **Hyperparameters**: γ is a key hyperparameter that weighs the relative importance of the inaccuracy cost versus the complexity cost in the calculation of PI. α is a sensitivity scaling parameter.

The value of PI ranges from 0 to 1. A system with a high PI value is considered to be able to efficiently use its WSI for synergistic information integration, thus making accurate predictions, reasonably assessing uncertainty, and integrating new information at a low cost.

However, consciousness is not only instantaneous but also continuous. To measure the **sustained strength and stability** of consciousness over a period of time, 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 penalizing the **volatility** 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 capability at every moment but also its predictive efficacy is stable and continuous. This aligns more closely with our intuitive understanding of a healthy, coherent, and awake state of consciousness.

3.3.2 From Ω to F: Why Maximizing Integration is the Optimal Strategy for Minimizing Free Energy

When proposing PI as a functional proxy for Ω_t, we must address a core theoretical challenge: could a "Clever Idiot" system exist—one that predicts the environment with high efficiency (high PI) but whose internal implementation is highly modular and lacks deep integration (low Ω_t)? IPWT's answer is no. We argue that in any complex cognitive system subject to real-world physical and computational constraints, **the close link between high PI and high Ω_t is not coincidental but a necessary consequence of the pressure for computational efficiency and model simplicity.** To formally prove this, we build an argument using first principles of information theory—specifically, the Minimum Description Length (MDL) principle.

The MDL principle reframes the problem of statistical inference as one of data compression. Its core idea is that the best generative model is the one that can describe the data with the shortest total length. According to Shannon's coding theory, the total length to describe the data, `L(Data)`, consists of two parts: `L(Data, Model) = L(Model) + L(Data | Model)`. Here, `L(Model)` is the number of bits required to describe the model itself (its structure and parameters), which directly quantifies the model's **complexity** `L(Data | Model)` is the number of bits required to describe the data given the model (typically the residuals between the data and the model's predictions), which directly quantifies the model's **goodness of fit** (proportional to the negative log-likelihood) [85], [109].

Based on this, we can ultimately prove that maximizing Ω is the optimal strategy for minimizing free energy (F) through a three-step derivation.

Argument 1 (Lemma 1): Minimizing free energy is computationally equivalent to seeking the minimum description length (MDL).
According to Friston's Free Energy Principle, the variational free energy `F` can be approximately decomposed into two core parts: `F ≈ Complexity - Accuracy`. We can establish a direct correspondence between the terms of `F` and MDL:

- **Model Complexity** corresponds to the number of bits required to describe the model itself, **`L(Model)`**. The more complex a model's structure, the more parameters it has, and the higher the required precision, the longer the bit string `L(Model)` needed to describe it.
- **Model Accuracy** corresponds to the negative of the number of bits required to describe the model's residuals, **`-L(Data | Model)`**. The higher the model's accuracy, the smaller the residuals between its predictions and the actual data. According to information theory, the number of bits `L(Data | Model)` required to describe a smaller, more predictable sequence of residuals is shorter. Therefore, `Accuracy ∝ -L(Data | Model)`.

Substituting these correspondences into the free energy formula, we get: `F ∝ L(Model) - (-L(Data | Model)) = L(Model) + L(Data | Model) = L(Total)`.

Conclusion: The process of minimizing variational free energy `F` is, on a computational and information-theoretic level, equivalent to finding a generative model that can describe the data with the shortest total length `L(Total)`.

Argument 2 (Lemma 2): A model's minimum description length `L(Model)` is functionally inversely proportional to its synergistic information integration (Ω).
To understand how `L(Model)` relates to synergistic information Ω, we must return to the formal definition of Partial Information Decomposition (PID). For two source variables X_1, X_2 and a target variable Y, the total mutual information is decomposed as: $I(X_1, X_2; Y) = \text{Red} + \text{Un}_1 + \text{Un}_2 + \text{Syn}$.

The length of `L(Model)` depends on the number of bits required to encode the model's internal generative rules (i.e., how to predict Y from X). Now let's compare two strategies:

- **Low-Ω Strategy**: A system with a low degree of synergistic information relies primarily on **redundant information (Red)** and **unique information (Un)**.
- Relying on `Red` means the model stores duplicate associations, e.g., "X_1 predicts Y alone" and "X_2 predicts Y alone."
- Relying on `Un` means the model stores independent, context-free rules, e.g., "if X_1 equals a, then Y equals c"; "if X_2 equals b, then Y equals d."

To explain the non-linear, higher-order interactions prevalent in the real world (e.g., Y equals z only when $X_1 = a$ AND $X_2 = b$), this strategy must rely on a huge, uncompressed look-up table or list of rules to exhaust all possible input combinations and their corresponding outputs. This inevitably leads to a very long model description length, `L(Model)`.

- **High-Ω Strategy**: A system with a high degree of synergistic information relies primarily on **synergistic information (Syn)**. By definition, synergy works by discovering and utilizing the irreducible, non-linear dependencies between variables that emerge only when considered as a whole. It uses a **single, compact rule** (e.g., a mathematical formula describing the interaction of X_1 and X_2) to capture this higher-order structure, rather than listing countless special cases. This is, in itself, an extremely efficient form of information compression. The number of bits required to encode this single generative rule is far less than that required for a massive look-up table.

Therefore, the higher a system's synergistic information integration `Ω`, the higher the compression ratio of its internal generative model, and the shorter its model description length `L(Model)`. We can formally express this functional inverse relationship as: `L(Model) ∝ 1/Ω`.

Conclusion: Synergistic information is a direct measure of the simplicity and compression efficiency of a model's internal rules. Maximizing synergistic information is equivalent to maximizing the model's compression ratio.

Argument 3 (Core Theorem): Maximizing synergistic information (Ω) is the optimal computational strategy for minimizing free energy (F).
From **Lemma 1**, we know that minimizing `F` is equivalent to minimizing `L(Total) = L(Model) + L(Data | Model)`. Assuming the model's accuracy is maintained (i.e., `L(Data | Model)` is relatively small), the main pressure to minimize `L(Total)` falls on minimizing `L(Model)`. From **Lemma 2**, we know that the optimal strategy for minimizing `L(Model)` is to maximize the model's synergistic information integration `Ω`.

Final Conclusion: The evolutionary and learning pressure to minimize free energy will inevitably drive a system to adopt computational strategies that maximize `Ω`. This is because high-`Ω` models are intrinsically simpler and more efficient; they can achieve the required predictive accuracy with the lowest complexity cost (shortest `L(Model)`), thereby achieving the lowest total free energy `F`.

Contemporary Large Language Models (LLMs) provide an excellent modern illustration of the "Clever Idiot." These systems can exhibit extremely high predictive integrity (high PI) on specific tasks, but they achieve this through pattern matching on massive amounts of data, lacking deep synergistic integration internally (low Ω_t). They perfectly expose the inherent fragility of such systems: enormous resource consumption, lack of autonomous agency, and rapid obsolescence in dynamically changing environments due to brittle generalization capabilities.

Therefore, the core argument of IPWT is that a cognitive system's ability to consistently exhibit high PI over the long term, stably, and across diverse contexts, in itself implies an intrinsic requirement for its internal information processing to be highly coherent, deeply integrated, and synergistically efficient (i.e., to have a high Ω_t).

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 dialogue with the empirical world will ultimately remain a castle in the air. Therefore, this chapter aims to outline the main neurobiological validation paths for IPWT, proposing a series of specific, operational experimental paradigms and predictions to transform 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 Macro-level Integration Metrics**: We will explore how to correlate IPWT's core computational metrics (especially PI/∫PI) with existing, widely validated macroscopic consciousness level indicators (such as the Perturbational Complexity Index, PCI) to establish the external validity of our theory.
2. **Evidence from Meso-level Network Dynamics**: We will delve into the level of functional networks and neural oscillations to explore what specific neuroimaging (e.g., fMRI) and electrophysiological (e.g., EEG/MEG) signal features might correspond to processes such as the dynamic formation of WSIs, information integration, and broadcasting.
3. **Testing through Micro-level Behavior and Psychophysics**: We can design a series of behavioral and psychophysical experiments to test specific predictions of IPWT regarding conscious access, attentional modulation, etc., by precisely manipulating subjects' perceptual and cognitive states.

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

4.1 Perturbational Complexity Index (PCI) and the 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 electroencephalographic (EEG) response [20]. This method has been validated in numerous clinical and experimental settings, reliably distinguishing between wakefulness, sleep, anesthesia, and different degrees of disorders of consciousness, and is considered one of the most reliable "consciousness-meters" available today [30], [110].

Within the IPWT framework, we believe there is a profound theoretical correspondence 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 relationship of **evocation-inference sampling**:

- **PCI measures the potential for integration**: PCI evokes 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 stimulation conditions, how complex an activity pattern can this brain network produce at most?

- **PI measures the efficiency of integration**: In contrast, PI does not externally perturb the brain. Instead, by modeling the brain's spontaneous data during **endogenous cognitive activity**, it **infers** the **information integration efficiency** actually achieved by its Workspace Instance (WSI) at a specific moment. It answers the question: In its current natural state, to what extent is this brain network engaged in effective prediction and information integration?

Thus, PCI is like stress-testing a car engine to find its maximum horsepower, while PI is like analyzing the trip computer data to infer the engine's actual fuel efficiency and operational smoothness during daily driving. A system with a powerful engine (high PCI) should also exhibit high efficiency during smooth driving (high PI).

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 a group of subjects spanning different levels of consciousness (e.g., from wakefulness to anesthesia, or among patients with different disorders of consciousness), the PCI values measured by TMS-EEG should show a significant positive correlation with the PI/∫PI values calculated from synchronously recorded resting-state neural data (such as fMRI or EEG). That is, a brain with higher integration potential should also exhibit higher predictive integrity at rest.
2. **PCI as a Gold Standard for Calibrating PI Parameters**: Since the calculation of PI involves some hyperparameters (like α, γ, δ) that need to be determined from empirical data, we predict that PCI can be used as an external criterion or physical anchor for calibrating these parameters. Specifically, we can adjust these hyperparameters to maximize the correlation between the calculated PI/∫PI values and the PCI values in the same group of subjects. This would tightly link the purely computation-based PI model with a widely accepted, physically-based physiological measurement, thereby greatly enhancing the empirical foundation of our theory. The recent study by Stikvoort et al. (2024), which found that the non-equilibrium dynamics of a whole-brain model can predict its perturbational complexity, also indirectly supports the feasibility of inferring perturbational responses from endogenous dynamics [111].

4.2 Neuroimaging Evidence

In addition to correlating with macro-level indicators, the core mechanisms of IPWT—such as the dynamic formation of WSIs, information integration, and broadcasting—should also have their neurophysiological imprints at the meso-scale of functional networks. We propose the following predictions that can be tested using neuroimaging (fMRI).

- **Dynamic Functional Networks and Connectomics of WSIs**: 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 of functional connectivity within and between specific brain networks during the execution of cognitive tasks that require conscious participation. These transiently enhanced functional networks are what we define as WSIs.

2. **Changes in Integration**: In the conscious awake state, networks associated with the dominant WSI (likely a combination of DMN and ECN, according to Luppi et al. [23]) will exhibit higher network integration (which can be quantified by graph theory metrics like global efficiency, or information-theoretic metrics like ΦR). When the level of consciousness decreases (e.g., during 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 provide background support for this prediction [112]–[115].

A recent landmark, large-scale adversarial collaboration study directly tested the conflicting predictions of GWT/GNWT and IIT, and its results fundamentally challenged the classic GNWT model [116]. This study, using multimodal neuroimaging techniques such as fMRI, MEG, and iEEG, systematically examined the neural representation of conscious content. Its core findings can be summarized in two points:

1. **Challenge to Content Representation - Incomplete Broadcast by the PFC**: A core prediction of GNWT is that any content entering subjective consciousness should be decodable from the PFC, as information needs to be globally broadcast through it. 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 could not represent the **fine-grained features** (e.g., the orientation of a face) that were also clearly perceived by the subjects. These fine-grained features could only be stably decoded from the posterior cortex (such as the occipital and parietal lobes). This finding directly challenges the completeness of the global broadcast, suggesting that the PFC may not be the broadcast center for all conscious content, but rather a selective amplifier that only broadcasts specific types of information (perhaps more abstract, more task-relevant).
2. **Challenge to Temporal Dynamics - The Missing Extinction Signal**: Another key prediction of GNWT is that the workspace updates its content through discrete ignition events. This means that when a conscious experience begins and **ends**, it should be accompanied by an ignition in the PFC. However, the experimental results clearly showed that while a strong activation (i.e., ignition) could be observed in the PFC at the **onset** of a stimulus, the expected ignition or extinction signal was **not** observed when the stimulus **disappeared** and the 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 constantly evolving. In their recent work, Baars et al. (2021) have attempted to develop GWT from a static, anatomically fixed model into a more dynamic and flexible **Global Workspace Dynamics (GWD)** framework [57]. They emphasize that conscious function is the result of the integration of the broad cortico-thalamic system, and its center of ignition is not necessarily in the PFC but can flexibly migrate within the cortical network according to task demands. This view partially 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 the IPWT theory. IPWT precisely carves out a more refined neuro-computational process of consciousness generation by functionally dividing the workspace into heterogeneous parts—namely, the DMN gateway responsible for synergistic information integration and the ECN broadcaster responsible for information distribution. The findings of Cogitate (2025), especially the "incompleteness" of PFC content representation and the "asymmetry" of its temporal dynamics, can be perfectly explained by the IPWT framework as the functional characteristics of the ECN as a broadcaster. It does not need (and should not) replicate all the integration details from the DMN gateway, but is only responsible for broadcasting the integrated, **decisional information** used to guide behavior. Therefore, this evidence, which poses a challenge to the classic GNWT, becomes strong support for the IPWT theory.

4.3 Behavioral and Psychophysical Experimental Design

In addition to validation at the macro and meso neurophysiological levels, carefully designed behavioral and psychophysical experiments can provide key evidence for the core predictions of IPWT from the level of individual subjective experience and behavioral performance. The advantage of such experiments is the ability to precisely manipulate stimuli and tasks and to obtain direct subjective reports from subjects, thereby building a bridge between computational models, neural activity, and phenomenal experience.

- **Perceptual Thresholds and Conscious Reports**: We can study the key factors determining whether a stimulus can enter subjective consciousness using classic paradigms that present stimuli near the perceptual threshold (e.g., visual masking, binocular rivalry) [117].

1. **Experimental Design**: In a visual masking task, a target stimulus (e.g., a letter) is obscured by a subsequent masking stimulus (e.g., a jumble of lines). By precisely adjusting the time interval between the target and the mask (Stimulus Onset Asynchrony, SOA), we can systematically manipulate the subject's conscious access to the target stimulus.
2. **IPWT Prediction**: We predict that a subject will only have a clear subjective conscious experience and be able to accurately report the stimulus when the information carried by the target stimulus reaches a sufficient degree of integration in the WSI (i.e., the calculated instantaneous PI value exceeds a certain threshold). At the threshold condition, the subject's reports will exhibit an all-or-none characteristic, corresponding to the non-linear dynamic process of whether the information successfully ignites and is broadcast in the WSI. We can test whether each successful report is accompanied by a significant late ERP component (such as P300) and a leap in the PI value by synchronously recording EEG/MEG data.

- **Attention and Multitasking**: Attention is the crucial spotlight that determines which information can enter the stage of consciousness [118]. We can study how attention affects the integration efficiency of the WSI by manipulating its allocation.

1. **Experimental Design**: A dual-task paradigm can be used, for example, requiring the subject 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), the cognitive resources for processing the second target (T2) will be temporarily depleted, causing the WSI for processing T2 to fail to form effectively or its integration efficiency (PI value) to drop sharply, thus preventing T2 from being consciously reported. Focused

attention will enhance the integration and PI value of the target WSI, while divided attention or high cognitive load may lead to a decrease or sharp fluctuation in the PI value, thereby impairing conscious perceptual performance.

- **Simulation of Specific Cognitive Dysfunctions**: In healthy subjects, non-invasive brain stimulation techniques (like TMS) can be used to briefly and reversibly simulate the core features of certain cognitive dysfunctions caused by brain damage, thereby studying their neuro-computational mechanisms under controlled conditions.

1. **Experimental Design**: For example, blindsight can be briefly simulated by applying inhibitory TMS to the primary visual cortex (V1). The subject is asked to report their visual experience in the affected visual field and simultaneously complete a forced-choice task on stimuli in that area.
2. **IPWT Prediction**: We predict that this inhibition of V1 will significantly reduce the quality of information flowing from this area to the dominant WSI, leading to a calculated PI value far below the conscious threshold, and the subject will report seeing nothing. However, the residual information may still be processed by local modules through other pathways (such as subcortical pathways), sufficient to support above-chance performance on the forced-choice task. This experiment would provide causal evidence for the IPWT explanation of blindsight as a failure of DWSI integration [119]–[122].

Finally, as a developing theoretical framework, when discussing the **falsifiability** of IPWT, we emphasize the openness and revisability of the theory. If any one or more of the above core predictions are reliably refuted by experimental evidence (e.g., if it is found that PI and PCI are uncorrelated at all levels of consciousness, or that the integration of the WSI is completely decoupled from subjective reports), then the IPWT theory itself must be significantly adjusted or even abandoned based on the new findings. This embrace of falsifiability is a necessary prerequisite for the development of any serious scientific theory.

5 Explanatory Power of IPWT for Conscious Phenomena: Analyzing Diversity within a Unified Framework

The power of a truly unified theory of consciousness lies not only in its mechanistic explanation of normal waking consciousness but also in its ability to provide a unified, internally consistent, and computationally principled framework for understanding those seemingly bizarre and perplexing special states of consciousness—including various abnormal subjective experiences caused by brain damage, mental illness, physiological changes, or drug effects. The core task of this chapter is to demonstrate how IPWT uses its key concepts, such as predictive coding abnormalities, Workspace Instance (WSI) dysfunction, changes in information integration efficiency (PI/Ω), and dysregulation of neural gating mechanisms, to systematically elucidate the neuro-computational basis of these states.

We no longer view these special conscious states as isolated oddities requiring separate theories for explanation. Instead, we see them as different system states arising from changes in specific parameters or modules within the unified computational space described by IPWT. For example, the latest research by Luppi et al. (2024) has already shown that the loss of consciousness under general anesthesia can be precisely described as a significant decline in the information integration capacity (ΦR) within the synergistic workspace (especially the DMN gateway) [23]. Following this line of thought, we will dissect a series of classic

consciousness puzzles, from blindsight and psychedelic states to schizophrenia and dissociative identity disorder, one by one, to show how IPWT provides new, deeper, and more integrated explanations for them.

5.1 Blindsight: Local Predictive Coding amidst DWSI Integration Failure

The phenomenon of blindsight, where patients with damage to the primary visual cortex (V1) report no subjective visual awareness in their blind visual field but can still perform above-chance on forced-choice tasks regarding visual stimuli (such as location, orientation) in that area [123], [124], presents a severe challenge to any theory of consciousness and provides a classic test case for IPWT. We reinterpret blindsight as a problem of the hierarchy and pathways of information integration.

In the IPWT framework, blindsight is not a single paradox but the result of two different information processing pathways coexisting but being dissociated from each other:

- **Integration Failure of the Dominant Workspace Instance (DWSI)**: In normal visual perception, information from the retina passes through the thalamus and first reaches V1 for initial processing. V1, as the key gateway for visual information to enter the entire cortical processing hierarchy, is crucial for forming high-definition, rich visual consciousness. When V1 is damaged, this main pathway to the DWSI responsible for generating subjective visual consciousness is severely disrupted. Due to severely deficient or extremely low-quality input information, the DWSI cannot construct an internal representation within the damaged visual field that has sufficiently low prediction error and high information integration. Therefore, the **Predictive Integrity (PI) of visual information from this area is extremely low** within the DWSI, failing to reach the integration threshold required to produce a subjective visual experience. This perfectly explains why patients insist they see nothing.
- **Residual Predictive Coding and Active Inference of Local Modules**: Although the main 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) [125]. These pathways connect to highly **specialized, unconscious modules** (e.g., those responsible for rapid motion detection or object localization). These local modules can still independently and locally conduct their own **predictive coding loops**. The visual input they receive is coarse but 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 minimization of these local prediction errors can directly drive **active inference and behavior** (e.g., saccading to a moving stimulus, or pointing to a light spot), completely without this information being integrated into the DWSI and producing subjective consciousness. This clearly explains why patients can still guess correctly, i.e., "see" behaviorally.

Therefore, IPWT's unique contribution is to clearly point out 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 to produce subjective experience**. It transforms blindsight from a seemingly mysterious philosophical paradox into a computationally addressable problem of information routing and integration efficiency within the cognitive

system. This explanation not only integrates known neuroanatomical evidence but also provides a theoretical possibility for future attempts to restore partial subjective visual experience through neuro-modulatory techniques (such as targeted stimulation of subcortical pathways).

5.2 Psychedelic States: Aberrant Enhancement of Prediction Errors 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 the blurring or even dissolution of the sense of self, provide a unique and controllable window for exploring the neuro-computational basis of consciousness [126], [127]. Within the IPWT framework, we trace these rich phenomenological features back to systematic changes in predictive coding parameters and workspace dynamic properties.

- **Aberrant Amplification of Prediction Error Signals and Degating of Information Flow**: We adopt and extend the REBUS (Relaxed Beliefs Under Psychedelics) and the Anarchic Brain model proposed by Carhart-Harris and Friston [128]–[130]. This model posits that the core action of classic psychedelics, as strong agonists of the serotonin 5-HT2A receptor, is to systematically **reduce or relax the precision-weighting of high-level predictions (priors)**. In the IPWT framework, this means that the inhibitory effect of top-down predictive signals 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, significant **prediction error signals**. These aberrantly amplified error signals flood the WSI, overwhelming normal top-down predictions, thus producing vivid perceptual distortions and audiovisual hallucinations. At the same time, the activation of 5-HT2A receptors may also weaken the **neural gating mechanisms** of the WSI, allowing information flows that are normally inhibited between different processing channels to freely enter the WSI and be abnormally integrated, providing a basis for subsequent phenomena like synesthesia. Recent research using network control theory corroborates this, finding that DMT significantly reduces the control energy required to drive brain state transitions, making the brain more accessible to different, typically hard-to-reach state spaces [131].
- **Blurring of WSI Boundaries and Formation of Temporary, Highly Integrated WSIs**: Under the dual influence of information flow degating and flattening of the predictive hierarchy, the functional boundaries between different WSIs, or between a WSI and originally independent specialized modules, may become blurred or even temporarily dissolve. This excellently explains the occurrence of **synesthesia**, such as hearing colors or seeing sounds, which can be understood as the abnormal and forced integration of auditory and visual information within the same WSI. More interestingly, we speculate that the system may even form some **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 producing those ineffable, profoundly meaningful experiences of cosmic unity or peak experiences, where PI and Ω values may reach extremely high levels for a short period.
- **Remodeling of the Self-Model 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). Under the influence of psychedelics, the WSI responsible for representing and maintaining the self-model undergoes deconstruction and remodeling because the interoceptive and exteroceptive prediction error signals it receives about the body, emotions, and memory are drastically altered, causing the predictive integrity (PI) of its original, stable self-representation to plummet. This perfectly explains the common subjective changes in psychedelic experiences, such as a blurred sense of self, a feeling of merging with the environment, and the most profound **ego dissolution** [132]. From a computational perspective, ego dissolution can be understood as the PI value of the self-model temporarily approaching zero, leading to a complete collapse of the boundary between self and non-self.

In summary, IPWT traces the rich phenomenological features of the psychedelic experience back to systematic changes in computable predictive processing parameters (like prior precision) and workspace dynamic properties (like gating mechanisms, boundary stability). This provides a new and guiding theoretical perspective for understanding the neural mechanisms of psychedelic drugs and their great potential in treating mental disorders such as depression and PTSD [133].

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

Schizophrenia, known for its complex spectrum of symptoms (including positive symptoms like hallucinations and delusions, as well as negative and cognitive symptoms), has long been a major challenge in psychiatry and neuroscience. Within the IPWT framework, we no longer view these seemingly heterogeneous symptoms as independent modular deficits but understand them uniformly as a **fundamental dysregulation of the predictive coding process**, and the resulting chain of **multiple functional impairments of the WSI in information integration, content gating, and boundary maintenance** [68].

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

1. **Hallucinations**: Auditory Verbal Hallucinations (AVH), in particular, can be understood as the system's internal model **spontaneously generating high-confidence (high precision-weighted) predictions**, for example, predicting that a voice will be heard. 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., there is a **failure in metacognitive-level monitoring of prediction errors** [134]. Consequently, these internally generated "voices" are erroneously attributed to the external world, leading to an irresistible and real experience of auditory hallucinations.
2. **Delusions**: Delusions can be seen as a **pathological belief system that is internally highly coherent** but severely detached from objective reality, which the system is forced to construct to explain persistent, abnormal perceptual experiences (i.e., persistent, high-intensity prediction errors) that cannot be understood by a normal world model. 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 aberrant predictions, the dysfunction of the WSI itself is central to the cognitive deficits in schizophrenia.

1. **Formal Thought Disorder**: Symptoms such as disorganized speech and incoherence of thought may directly reflect the WSI's impaired ability to effectively select, organize, and integrate information to form a coherent stream of thought. This may be related to a **dysregulation of the WSI's neural gating mechanisms**, which prevents it from effectively filtering out irrelevant internal and external information and from maintaining a stable, goal-directed cognitive state.
2. **Passivity Phenomena**: Bizarre symptoms of self-boundary disturbance, such as thought insertion and the feeling 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-produced content may have abnormally increased, or the neural tags used to label self-related information (possibly based on interoceptive prediction) have malfunctioned. This leads to the patient's inability to accurately distinguish which thoughts and actions originate from their own intentions and which come from external forces, resulting in the uncanny experience of being controlled by an outside power.
3. **Abnormal Network Connectivity**: The latest research in network neuroscience 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 lower global communication efficiency (functional integration) [135], [136]. This is consistent with IPWT's view that both the integration and broadcast functions of the WSI are compromised.

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., abnormalities in 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 (known as "alters") within a single individual. These identity states recurrently take control of the individual's behavior and are accompanied by extensive memory gaps that cannot be explained by ordinary forgetfulness. Phenomenologically, DID appears to be the most extreme challenge to the unity of consciousness. IPWT offers a feasible computational model for this seemingly mysterious phenomenon, based on the dynamic switching and information segregation of Workspace Instances (WSIs).

We do not believe that DID patients have 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 have failed to develop a single, unified internal generative model and WSI. Instead, it has formed **multiple, potentially independent WSI systems**.

- **Multiple Potential WSI Systems**: We hypothesize that each "alter" corresponds to a relatively independent WSI system. Each WSI system is associated with a unique set of internal generative models, which contain the specific memories, beliefs, behavioral patterns, and emotional response tendencies of that alter. For example, the WSI of a protective alter might primarily consist of predictive models related to threat response, while the WSI of a child alter might be dominated by models related to early memories and attachment. Neuroimaging studies provide corroborating evidence, finding significant and replicable differences in brain activity patterns and functional connectivity in DID patients across different identity states [137]–[140].
- **Dynamic Flip of Dominant Workspace Instance (DWSI) Status**: In the IPWT framework, at any given moment, typically only one WSI system can occupy the **dominant position (Dominant WSI, DWSI)**, and its content and processing constitute the individual's current conscious experience and overt behavior. We believe that the dramatic personality switching in DID can be precisely understood as a **dynamic, and often rapid, flip of the DWSI status among these potential WSI systems**. This flip can be triggered by specific external cues (such as trauma-related reminders) or changes in internal state, causing a previously background WSI system to become activated and replace the currently dominant one.
- **Neural Gating Mechanisms and Information Segregation**: When a particular WSI system becomes the DWSI, other non-dominant WSI systems are largely suppressed or functionally segregated by powerful **neural gating mechanisms**. This segregation mechanism is key to the common **amnesia** found in DID. When alter A is dominant, their experiences and learned information are integrated into A's WSI system. When the system flips to alter B's dominance, B will not be able to easily access A's memory content due to the functional segregation between B's WSI and A's WSI, thus manifesting as amnesia for events that occurred during A's period of dominance. A study by Reinders et al. (2019) using pattern recognition methods to analyze brain structural images of DID patients found that they could be distinguished from healthy individuals based on morphological features of the brain, providing further evidence for a neurobiological basis of DID [141]–[145].

Thus, IPWT transforms DID from an incomprehensible psychological phenomenon into a computational problem of 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., psychotherapy 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 Systems

The core feature of Depersonalization/Derealization Disorder is a persistent or recurrent experience of detachment from, and unreality about, oneself (depersonalization) or one's surroundings (derealization), as if in a dream. The patient's reality testing is usually intact; they know the feeling is not real, but they cannot shake the intense sense of alienation. In the IPWT framework, we argue that this disorder does not stem from a collapse or derangement of the WSI itself, but from a **significant weakening or abnormal alteration of the functional**

connectivity between the Dominant Workspace Instance (DWSI) and the modules responsible for processing specific sensory information or imbuing experience with emotional color.

- **Depersonalization: Reduced Precision-Weighting of Self-Related Interoceptive Predictions** The core experience of depersonalization is feeling unreal or like an outside observer of one's own thoughts, feelings, or body. We interpret this as a **significant reduction in the effective connectivity between the DWSI and the interoceptive and proprioceptive input streams it normally receives from the somatosensory, limbic, and episodic memory systems**. From a predictive coding perspective, this means that the system assigns **abnormally low precision-weighting** to the prediction error signals coming from one's own body and emotional states. Therefore, although the body's physiological activities continue, these signals cannot effectively update the self-model in the DWSI. The DWSI loses its tight connection with the body's here-and-now feelings, causing the subjective experience to lose its sense of personal ownership, belonging, and realness, as if an invisible glass wall separates the conscious self from the physical existence of the body.
- **Derealization: Dissociation of External Sensory Information and Affective Tagging** The core experience of derealization is feeling that the external world is unreal, hazy, or like being in a dream or a movie [146], [147]. We interpret this as, although external sensory information (like vision, hearing) can be processed normally and enter the DWSI, the DWSI fails to **effectively and synchronously connect with the emotional appraisal modules (like the amygdala) or contextual association modules (like the hippocampus)** that are responsible for imbuing it with meaning and emotional salience during integration. This results in the external world being clearly perceived but appearing flat, distant, and meaningless in subjective experience. It is akin to a functional emotional blindness, where sensory information loses its usual accompanying affective tag. This affective tagging is crucial for our sense of the world as real and personally relevant. 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. The adaptive gain theory proposed by Aston-Jones & Cohen (2005) in their study of the norepinephrine system also emphasizes how the brain dynamically adjusts its processing of information based on task relevance, which is analogous to our proposed role of emotional tagging in conferring a sense of reality [148].

Thus, IPWT transforms depersonalization/derealization disorder from a vague psychological description into a testable neuro-computational hypothesis about connectivity abnormalities between the DWSI and specific functional modules and dysregulation of prediction signal precision-weighting.

5.6 Lucid Dreaming: Parallel WSIs and the Neural 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 content of the dream, provides a fascinating natural experiment for studying the hierarchical structure of consciousness, the parallel operation of WSIs, and metacognitive functions [149], [150]. Within the IPWT framework, lucid dreaming can be understood as a special, hybrid state of consciousness in which at least two functionally different WSI systems are abnormally co-activated and interact with each other.

- **Parallel Activation of a Dream WSI and a 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 coherent. This WSI is responsible for generating the vivid, bizarre dream content we experience. In lucid dreams, we hypothesize that in addition to this dream WSI, another **metacognitive WSI**, which is typically active only in the waking state and is closely associated with self-awareness and reality testing functions (possibly overlapping with parts of the prefrontal cortex), is abnormally co-activated. Therefore, lucid dreaming is a hybrid state where two WSIs operate in parallel: one is acting in the play, while the other is watching it.
- **Recognition of Higher-Order Prediction Errors and the Emergence of Lucidity**: How does the sense of lucidity arise? We believe it stems from the metacognitive WSI identifying **higher-order prediction errors** while monitoring the content presented by the dream WSI. The metacognitive WSI possesses higher-level, more stable predictive models about the world and one's own abilities (e.g., "humans cannot fly," "I should be lying in bed sleeping right now"). When it detects 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 massive higher-order prediction error. To minimize this error, the metacognitive WSI generates a new, best inference to explain this mismatch, and that inference is: **I am dreaming**. The generation of this inference, at the phenomenological level, corresponds to the emergence of the sense of lucidity.
- **Limited Interaction between WSIs and Dream Control**: Once lucidity is achieved, a degree of bidirectional information exchange may be established between the metacognitive WSI and the dream WSI. This interaction gives the dreamer the ability to control the dream. From a computational perspective, this can be understood as the metacognitive WSI beginning to actively influence or hijack the content generation process of the dream WSI through top-down predictive signals. For example, a lucid dreamer can change the dream scene or summon specific characters by intention, 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 ability to control the dream is often limited and unstable, which may reflect that the information interaction bandwidth between the two WSI systems is limited and that a degree of functional segregation still exists between them.

Thus, IPWT decomposes the strange experience of lucid dreaming into a series of researchable computational processes, such as the parallel activation of WSIs, the monitoring of higher-order prediction errors, and the information interaction between WSIs, providing clear theoretical guidance for future research into, and even induction of, lucid dreams using neuroimaging or neuromodulation techniques.

5.7 Locked-in Syndrome: Intact Integration but Failed Broadcast of Consciousness

Locked-in Syndrome (LIS) is a rare but devastating neurological disorder in which patients, typically due to a lesion in the ventral pons, suffer from almost complete motor paralysis, including speech and limb movements, while their consciousness, cognitive functions, and vertical eye movements (or blinking) are usually intact. This state of being a conscious prisoner

provides the most direct and compelling clinical evidence for the functional dissociation between information integration and information broadcast in the IPWT framework.

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

- **Intact Integration Function of the Dominant Workspace Instance (DWSI)**: The cerebral cortex of LIS patients, especially the Default Mode Network (DMN) as a gateway region responsible for high-level cognitive functions and information integration, is usually structurally and functionally intact. This means that patients can normally receive, process, and integrate various streams of information from internal (e.g., thoughts, emotions, memories) and external (e.g., hearing, vision, although external input may be limited) sources. Their DWSI can construct coherent, unified models of the world and the self, and thus their **instantaneous information integration (Ω_t) and predictive integrity (PI) are at high levels**, supporting a full, waking subjective conscious experience. The patient's inner world is rich and coherent; they can think, feel, remember, and understand.
- **Interruption of the Broadcaster's (ECN) Output Pathway**: However, the lesion in the ventral pons (usually an infarction or hemorrhage) precisely severs the descending motor pathways from the cerebral cortex (especially the Executive Control Network ECN as the broadcaster) to the spinal cord and peripheral muscles. This means that although the DWSI has successfully integrated action intentions and decisions (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. The information is "locked in" inside the conscious workspace, unable to be translated into observable overt behavior. Patients can only communicate in a limited way through residual vertical eye movements or blinking (these pathways are usually unaffected by ventral pons lesions), which further confirms the integrity of their internal consciousness [151].

Therefore, Locked-in Syndrome provides key **double dissociation** evidence for the functional division between the DMN gateway (responsible for integration) and the ECN broadcaster (responsible for distribution) in the IPWT theory. It forms a sharp contrast with blindsight (where information cannot effectively enter the DWSI for integration, leading to unconscious perception but with residual behavior), and together they reveal the different computational stages on which consciousness generation and behavioral output depend. Locked-in Syndrome emphasizes that even if information is perfectly integrated in the WSI and produces subjective consciousness, this consciousness cannot be known to or influence the external world without an effective broadcast mechanism. 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 that could bypass damaged pathways and directly read conscious content and intentions from the patient's DWSI.

6 Neuro-computational Reconstruction of Subjective Experience: Qualia as the Geometry of Inference Space

Any complete theory of consciousness must ultimately provide a computationally operational and neurobiologically plausible implementation for its core phenomenological claims. For the essence of subjective experience—Qualia—the "hard problem" [2], IPWT 2.0 proposes a fundamental reconstruction, deepening the previous version's functional labeling account into a more formal **geometric-dynamic theory** based on the Free Energy Principle (FEP) and

synergistic information (Ω). We argue that Qualia space is not a mysterious realm separate from the physical world; it is computationally **identical** to the **Inference Space** of the system's internal generative model. The geometry of this space is defined by the **synergistic information (Ω)** that the system can distinguish and integrate, while the "what-it-is-likeness" of subjective experience is the **dynamic process of active inference** as the system's state (posterior beliefs) traverses geodesics of free energy minimization in this high-dimensional inference space.

6.1 From IIT's Conceptual Structure to FEP's Inference Space

IIT's profound insight lies in recognizing that the structure of conscious experience (Qualia Space) is constituted by a "conceptual structure" (MICS) formed by the "concepts" the system can distinguish [24]. However, its definition of a "concept" relies on perturbational analysis of the system's physical causal structure, which is not only computationally intractable but, more importantly, fails to elucidate the specific role of this structure in cognitive function. IPWT retains its core insight—that the space of conscious experience is constituted by states the system can **distinguish**—but proposes that for any system adaptively surviving in its environment (i.e., a system obeying the Free Energy Principle), its functionally distinguishable state space is necessarily identical to its internal generative model's **inference space**. We establish this equivalence through the following argument.

Definition 1: IIT's "Concept" and FEP's "Posterior Belief."

- In IIT, a **Concept** is defined as a **maximally irreducible causal structure** (MICS) within the system. It specifies the causal effects this structure can have on its own past and future states from its own perspective. This is essentially a static, ontological description of the system's **intrinsic causal topology**.
- In FEP, a **Posterior Belief** is a probability distribution `p(s|o)` over the world's hidden states `s`, given sensory evidence `o`. It represents the system's **dynamic inference** about the causes of its sensory data. This is essentially a dynamic, epistemological description of the system's **functional knowledge state**.

Lemma 1: For a system following FEP, its existence is equivalent to its inference. According to FEP, any self-organizing system must resist entropy by minimizing its variational free energy. The core of this process is updating its posterior beliefs `p(s|o)` to optimize its predictions and explanations of the world. Therefore, what a system functionally "is" is entirely defined by what it "believes." The set of all possible, functionally relevant internal states a system can enter is the set of all possible posterior beliefs it can form. Any state that can never be inferred is, for the system, functionally non-existent.

Core Theorem: The inference space is the necessary and sufficient implementation of the conceptual structure. Based on the above lemma, we can demonstrate that FEP's Inference Space is a computationally feasible and functionally explicit implementation of IIT's Conceptual Structure.

- **Necessity**: An IIT concept, no matter how complex its underlying physical causal structure, is meaningless for the system's adaptive behavior if it can never become a posterior belief formed in the process of minimizing free energy during interaction with the environment. Therefore, a functionally meaningful concept **must** be mappable to one or a series of belief states in the inference space.

- **Sufficiency**: FEP's inference space satisfies all of IIT's core axiomatic requirements for the structure of conscious experience, but places them on a functional and computational foundation. **Intrinsic Existence** is embodied in the fact that inference is an operation of the system's internal generative model; **Information** is embodied in that each posterior belief specifies a particular probability distribution from a space of possibilities; **Integration** and **Exclusion** are realized through the geometry of the inference space shaped by synergistic information (Ω) and the dynamics of inferential competition that occurs within it, as we will elaborate later.

Conclusion: Thus, we complete a fundamental replacement. IPWT reframes IIT's static, ontological "conceptual structure" based on physical causality into FEP's dynamic, functional "inference space" based on Bayesian inference. Mathematically, the inference space is the space constituted by the probability distributions of all hidden variables in the system's generative model. Each "point" in this space represents a possible "belief" or "hypothesis" the system has about the state of the world (including itself) [63], laying the groundwork for our subsequent analysis of the structure of qualia using information geometry.

6.2 Synergistic Information (Ω) as the Geometric Constructor of Inference Space

Having established the inference space as the foundation of qualia space, we must answer a core question: what gives this space its structure, allowing for "experiential" differences between different "beliefs"? IPWT's answer is **synergistic information (Ω)**. We elucidate how synergistic information, by defining the information geometry of the inference space, becomes the constructor of the structure of qualia through the following argument.

Lemma 2.1: The dimensionality of the inference space is determined by statistically independent hidden variables. In a generative model `p(s_1, s_2, ..., s_n)`, if its hidden variables are statistically independent, i.e., `p(s_1, ..., s_n) = p(s_1)p(s_2)...p(s_n)`, then its inference space is topologically the Cartesian product of the individual subspaces `S_i`: `S_1 × S_2 × ... × S_n`. The geometry of this space is flat (Euclidean), as movement (belief updating) in one dimension does not affect the others.

Lemma 2.2: Synergistic information is a direct measure of the irreducible statistical dependencies in the posterior belief. When faced with sensory evidence `o`, the system forms a posterior belief `p(s_1, ..., s_n | o)`. If statistical dependencies arise among the hidden variables in this posterior belief, i.e., `p(s_1, ..., s_n | o) != p(s_1|o)p(s_2|o)...p(s_n|o)`, this dependency can be quantified through Partial Information Decomposition (PID). Here, **synergistic information** `Syn(s_1, ..., s_n; o)` precisely captures the part of the information that emerges only when all variables are considered as a whole. Therefore, the existence of synergistic information directly implies that the posterior belief is irreducible.

Core Theorem 2: Synergistic information (Ω) defines the non-Euclidean geometry of the inference space by imposing irreducible statistical constraints. In the framework of information geometry, the statistical dependencies in the posterior belief are equivalent to imposing a non-trivial metric tensor, the Fisher Information Matrix, on the manifold of the inference space. This transforms the inference space from a flat Euclidean space into a **curved Riemannian manifold**.

- **Irreducible Geometric Shapes**: The higher the synergistic information (Ω), the stronger the dependencies between variables, and the greater the "curvature" of the inference

space. This irreducible, high-dimensional geometric shape (manifold) defined by synergistic information is the structural basis of a quale (e.g., "a square").

- **Distance and Similarity of Qualia**: In this curved space, the distance between two qualia (i.e., two posterior beliefs `p_1` and `p_2`) is no longer the simple Euclidean distance but the **geodesic distance** defined by the KL divergence [34]. It measures how much "cost" (i.e., increase in free energy) the system needs to incur to move its belief from one state to another along the shortest path on the manifold. This provides a principled, quantifiable definition for the similarity of qualia.

6.3 The "What-it-is-likeness" of Qualia: Active Inference Dynamics in Inference Space

Having defined the geometry of qualia space, we further argue that the "what-it-is-likeness" and valence of subjective experience are the **dynamic process** of the system's posterior belief undergoing active inference within this geometric space. This forms a "Beautiful Loop" of cognition and phenomenal experience, where inferential competition, Bayesian binding, and epistemic depth co-emerge to give rise to conscious experience [94].

Lemma 3.1: Variational free energy F defines a potential energy landscape in the inference space. For any given posterior belief `q(s)`, the variational free energy `F(q)` can be formally defined as `F(q) = D_KL[q(s) || p(s)] - E_q[log p(o|s)]`, where `p(s)` is the prior belief and `p(o|s)` is the likelihood. For a given prior and sensory evidence `o`, `F` is a functional of the belief `q`. Therefore, free energy `F` defines a **potential energy landscape** over the inference space constituted by all possible posterior beliefs.

Core Theorem 3: The "what-it-is-likeness" and valence of subjective experience is the dynamic process of the posterior belief updating along the negative gradient of free energy in the inference space. According to the principle of active inference, the process of perceptual inference is mathematically equivalent to a **gradient descent** on the free energy landscape, with its dynamics described by the equation `∂q/∂t = -∇F(q)`. We propose that it is this dynamic process itself that constitutes the "what-it-is-likeness" of qualia.

- **What-it-is-likeness as a Dynamic Trajectory**: The "feeling" of a quale, such as the sensation of red, is the **dynamic trajectory** of the posterior belief falling into and stabilizing within an attractor region associated with "long-wavelength light" in the "color" subspace of the inference space.
- **Valence as a Property of the Gradient Field**: The valence of a quale, its intrinsic pleasant or unpleasant character, is determined by the **properties of the gradient field** along this dynamic trajectory.
  - **Negative valence (e.g., pain, fear)** corresponds to a steep gradient field far from a stable point. The system's state is pushed into a high free-energy region by a large, high-precision prediction error (such as a tissue damage signal) and experiences a very strong potential gradient pointing towards "escape." This strong, directed "repulsive force" is the computational basis of its negative valence.
  - **Positive valence (e.g., pleasure, satisfaction)** corresponds to a gentle gradient field approaching a stable attractor (a low free-energy region). The system's state finds a (temporary) optimal solution here, experiencing a very small or vanishing free-energy gradient, thus generating a feeling of "ease" or "contentment."

In this way, IPWT provides a unified, first-principles computational explanation for the three core aspects of qualia: structure, similarity, and what-it-is-likeness. On this basis, the **Functional Labeling** account from the previous version can be seen as a high-level, applied interpretation of this more fundamental geometric-dynamic theory. Qualia can play the role of efficient functional labels (e.g., driving behavior through information compression and salience tagging [152]) precisely because they are rooted in the geometric structure and dynamic processes of the system's inference space.

6.4 Quantifiable Dimensions and Neural Manipulability of Qualia

If Qualia are indeed the geometric and dynamic states of the system's inference space, this should generate a series of testable scientific predictions. In principle, we should be able to find measurable **computational parameters or neurodynamic metrics** that correspond to the different properties of Qualia (such as intensity, clarity, richness, valence, etc.). In other words, the quality of subjective experience should be mappable to the quantity of a computational model.

We propose that the following 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 the prediction error** processed by the WSI. A high-precision, large prediction error (i.e., high surprise) would produce a more intense, more attention-grabbing subjective experience.
- **Richness/Uniqueness of Experience**: May be related to the **complexity and synergy** (i.e., Ω_t or its proxy ΦR) of the information representation in the WSI. A highly integrated WSI state containing a large amount of synergistic information would correspond to a richer, more unique phenomenal experience. The Quality Space Computations framework proposed by Fleming & Shea (2024) is highly compatible with this idea; they also argue that the structure of subjective experience can be mapped onto a computable, high-dimensional representational space [153], [93].
- **Clarity/Durability of Experience**: May be related to the **stability and duration** (i.e., $\int$PI) of the WSI state. A stable, sustained WSI state with low volatility would correspond to a clearer, more stable subjective experience; conversely, a rapidly changing, unstable WSI state might correspond to a vague, confused experience.
- **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 produce feelings of confusion, discomfort, or even pain.

These proposed correspondences open up new, operationalizable paths for the study of Qualia. We can explore and validate them through several engineering approaches:

1. **Correlational Studies with Neuroimaging and Psychophysics**: Through precise psychophysical experiments (such as the visual masking and attentional blink paradigms mentioned earlier) to precisely manipulate subjects' subjective experiences, while synchronously recording their high-temporal-resolution neural activity (e.g., EEG/MEG). We can then test whether subjects' subjective reports on the intensity, clarity, etc., of their experience show significant correlations with specific parameters we calculate from the neural

data (such as the amplitude of prediction error signals, proxy measures of Ω_t, stability of the WSI, etc.).

2. **Direct Manipulation via Neural Interface Technologies**: With the development of neural interface technologies, we may in the future be able to directly and precisely manipulate the neurodynamic parameters of a specific WSI using techniques like TMS, focused ultrasound, or deep brain stimulation. For example, we could try to enhance or weaken the Gamma-band oscillations of a WSI through stimulation and then observe whether this systematically alters the subject's subjective report of the richness or clarity of their experience of the related stimulus.
3. **Replicating Functional Equivalents in Advanced AI**: We can attempt to implement functional equivalents of Qualia in artificial intelligence agents built on the IPWT architecture. For example, we could design an AI that, when its prediction error exceeds a certain threshold, enters a special alert state where all its computational resources are forcibly allocated to processing this error, and all its subsequent decisions are influenced by this alert. By studying the behavior and internal states of such an AI, we can better understand the functional role that Qualia might play in a purely computational system.

By transforming the Qualia problem from a purely ontological "what-it-is" question to a functional and mechanistic "what-it-does" and "how-it-is-realized" question, IPWT aims to pull it from the realm of philosophical speculation into a new framework where it can be investigated 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 integrative, computational, and verifiable theoretical framework. Having detailed the theoretical core of IPWT, its computational metrics, its explanatory power for various conscious phenomena, and its neurobiological validation paths, we will now take a step back in this chapter to conduct a macroscopic and critical discussion and outlook on IPWT's main theoretical contributions, the potential challenges and limitations it faces as an emerging theory, and the future research landscape it may open up for related fields such as consciousness science, clinical neuroscience, and artificial intelligence.

7.1 Main Theoretical Contributions and Core Advantages of IPWT

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

1. **Theoretical Integration and Coherence**: IPWT's primary contribution lies in its successful fusion of the three most important yet seemingly separate theoretical pillars of contemporary consciousness science—the dynamic mechanisms of PCT/FEP, the architectural functions of WT, and the phenomenological insights of IIT—into a unified and internally consistent theoretical framework. It is not a simple patchwork of these theories, but a creative functional reconstruction that reveals their deep connections: conscious content

is **generated** by predictive mechanisms, **integrated** in a 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 (∫PI) as computable proxy metrics, IPWT largely overcomes the computational complexity bottleneck faced by any theory based on information decomposition when applied to large-scale systems. This makes IPWT not just a philosophical framework for speculation, but a scientific tool that can be practically applied to analyze real neural data, fit models, and test hypotheses.

3. **Substrate Independence and Implications for AI**: Because IPWT defines the core mechanisms of consciousness at the level of information processing and computational function (such as prediction, integration, broadcast), it naturally supports the view of multiple realizability or substrate independence of consciousness. This means that consciousness is not a patent of the biological brain; any system that can implement the same functional architecture and information dynamics could, in principle, give rise to consciousness. This stance not only resolves the theoretical dilemma of IIT on this issue but also provides important theoretical guidance and design principles for the future construction of Artificial General Intelligence (AGI) systems with higher cognitive abilities and even some form of artificial consciousness [154]–[156].

4. **Unified Explanatory Power for Special States of Consciousness**: As detailed in Chapter 5, IPWT can provide a unified explanatory framework, based on common neuro-computational principles, for a wide range of normal, special, and pathological states of consciousness, 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 great explanatory breadth and depth of the theory, and providing new ideas for developing new diagnostic and therapeutic strategies in clinical neuroscience and psychiatry.

5. **Functional Account of the Qualia Problem**: IPWT bypasses the endless debate on the ontological status of Qualia. By redefining it as a functional label for the system's internal states, it provides a new perspective for this hard problem that can be investigated by scientific methods. This functionalist account emphasizes the key adaptive role of Qualia in information compression, salience tagging, and behavioral drive, and points to research paths that link its different dimensions (such as intensity, richness) with quantifiable computational parameters, thus moving Qualia research in a more operational and falsifiable direction.

7.2 Potential Challenges for IPWT and Future Directions for Neurobiological Research

Although IPWT exhibits numerous theoretical advantages and great explanatory potential, as an emerging theoretical framework, it inevitably faces a series of challenges that need further research and resolution. These challenges are not only the driving force for theoretical refinement but also point the way for future neuroscience research.

1. **Operationalization and Measurement Challenges of Core Concepts**: Although we have proposed PI/∫PI as a proxy for Ω_t and incorporated the latest advances in ΦR, the **precise calculation, robustness, and validity** of these metrics in real, complex neural data (i.e., the extent to which 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 time scales, are still active research questions in computational information theory. At the same time, precisely and in real-time identifying and delineating the boundaries of one or more WSIs, their constituent neuronal populations, and the functional connectivity patterns between them in a real, dynamically changing brain remains a huge technical challenge. Future research needs to develop more advanced computational tools and analytical methods to overcome these quantitative hurdles.
2. **Further Validation of the Neurobiological Basis**: Although each component of IPWT (predictive coding, workspace, information integration) has some basis in neuroscience research, after integrating them into a unified framework, its overall neural implementation mechanism still requires 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 how do they dissipate after a task is completed? What specific neural circuits, synaptic plasticity mechanisms, and neuromodulatory systems (such as acetylcholine, norepinephrine) are involved?
- **Specific Mechanisms of Information Integration and Broadcast**: How is synergistic information integrated within the WSI? What are the specific neuro-computational mechanisms of the DMN as a gateway and the ECN as a broadcaster? This may involve specific neural oscillation patterns (such as 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 dominance? What neural circuit gating and inhibition mechanisms are involved? Future research needs to combine multimodal neuroimaging (fMRI, EEG, MEG), electrophysiological recordings (in vivo/in vitro), optogenetics, and chemogenetics to conduct more refined causal experiments in animal models and human subjects.

1. **Balancing Model Complexity and Interpretability**: IPWT is a grand and multi-layered theory, and its corresponding computational models are very complex. How to maintain its interpretability and simplicity while pursuing the model's explanatory power and predictive accuracy, avoiding the model becoming too much of a black box, is an important challenge. We need to develop Explainable AI (XAI) methods that can reveal the internal workings of the model, so that the theory's predictions are not only accurate but also provide intuitive biological insights [157], [158].
2. **Deeper Exploration of the Qualia Problem**: Although the functional labeling view proposed by IPWT offers a new perspective for scientific inquiry into the Qualia problem, the extent to which it can truly touch the essence of subjective feeling remains an open philosophical and scientific question. Future research needs to more deeply explore how the multidimensional properties of Qualia are precisely mapped onto neuro-computational parameters and to try to establish causal relationships for these mappings through more precise psychophysical experiments and neuromodulation.

3. **Integration with Other Cognitive Functions (e.g., Emotion, Motivation, Social Cognition)**: Consciousness does not exist in isolation; it is closely 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 more deeply integrate these important cognitive dimensions. For example, how to link the predictive coding of emotions (such as interoceptive prediction) with the functional labeling of Qualia, and how social interactions shape an individual's world model and self-model. This will help to build a more comprehensive and ecologically valid model of the human mind.

7.3 Future Research Directions and Potential Impact of IPWT

Faced with the above challenges, the future development of IPWT needs to proceed synergistically on multiple levels, including theoretical deepening, computational modeling, experimental validation, and clinical application. We believe that through interdisciplinary fusion, IPWT has the potential to have a profound impact in the following key areas.

1. **Mathematical and Formal Deepening of the Theory**: The core of IPWT lies in its information-theoretic foundation. Future research needs to continue exploring the mathematical frontiers of PID and ΦID theory, especially how to develop more efficient and robust methods for calculating synergistic information, enabling its application to larger-scale, more complex neural system data. For example, computational innovations like the Fast Möbius Transform proposed by Jansma et al. (2024) offer new possibilities for applying ΦID in large systems [159]. At the same time, it is necessary to more closely integrate the mathematical framework of FEP with the dynamic organization and information integration processes of the WSI. For example, how to derive the necessity of information integration within the WSI from the variational free energy minimization of FEP, and how to formalize the self-organized criticality of the WSI.
2. **Construction and Validation of Multi-scale, Multimodal Neuro-computational Models**: IPWT requires more refined computational models that are constrained by neurobiology. Future research should focus on:

- **Multi-scale Modeling**: Building IPWT models that can bridge microscopic neuronal activity (e.g., spike firing, synaptic plasticity) with macroscopic brain region activity (e.g., fMRI BOLD signals, EEG oscillations) to validate theoretical predictions at different spatial and temporal scales.
- **Multimodal Data Integration**: Using multimodal neuroimaging data (e.g., fMRI, EEG/MEG, DTI, PET) and machine learning techniques to fit IPWT model parameters at the individual level, exploring their associations with cognitive functions, personality traits, and mental health. For example, combining the spatial resolution of fMRI with the temporal resolution of EEG/MEG to dynamically track the formation and information flow of WSIs.

1. **Translational Research and Clinical Applications for Special States of Consciousness**: IPWT provides a unified computational pathology 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 computational models, such as monitoring abnormal patterns of PI/∫PI or ΦR to assist in the diagnosis of schizophrenia, DID, or to assess the level of consciousness in patients with disorders of consciousness.

- **Precision Intervention Protocols**: Designing cognitive training, psychotherapy, or neuromodulation intervention protocols aimed at repairing specific computational stages. For example, using targeted stimulation (e.g., rTMS, DBS) to modulate the neural gating mechanisms of the WSI or to enhance its information integration efficiency, thereby improving patients' symptoms. The findings of Luppi et al. (2024) on the changes in ΦR during anesthesia provide an empirical basis for such integration-based interventions [23], [160].

1. **Exploratory Applications in Artificial Intelligence and Brain-inspired Computing**: IPWT's substrate independence has important implications for the field of artificial intelligence. Future research could:

- **Design New-generation AGI Architectures**: Draw on the core principles of IPWT (predictive coding, WSIs, synergistic information integration) to design Artificial General Intelligence (AGI) systems with greater autonomous learning, generalization, and adaptation capabilities. For example, building neural network architectures with hierarchical predictive models and dynamic workspaces [161]–[163].
- **Explore the Implementation of Functional Labels**: Explore the implementation of functional equivalents of Qualia in AI systems, i.e., mechanisms that can form high-order representations of their own internal states and assign behavioral value to them. This will help us understand how consciousness might arise in artificial systems and could potentially endow AI systems with stronger metacognitive abilities and a nascent form of self-awareness.

1. **In-depth Exploration of Philosophical and Ethical Dimensions**: The development of IPWT will inevitably raise profound philosophical and ethical questions about the moral status of artificial consciousness, the understanding of free will, and the ethical boundaries of neuromodulation technologies. Future research will require interdisciplinary dialogue with philosophers, ethicists, and sociologists to jointly explore the societal impacts of these cutting-edge scientific developments.

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

The Integrated Predictive Workspace Theory (IPWT), by deeply fusing 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, mechanism, and function of consciousness. IPWT views consciousness as an emergent property of a dynamic information processing process, driven by prediction, centered on the logical irreducibility of information integration (synergistic information), and aimed at minimizing free energy, all occurring within a specific functional architecture (the Workspace Instance, WSI). The core contributions of IPWT are:

1. **Theoretical Integration**: It successfully merges the strengths of the three major mainstream theories of consciousness into a more comprehensive and coherent perspective.
2. **Computational Feasibility**: By introducing computable proxy metrics (such as Predictive Integrity PI and its integral ∫PI, as well as the revised Φ value ΦR), IPWT moves consciousness research from philosophical speculation and qualitative description towards 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 a wide range of special states of consciousness (such as blindsight, psychedelic states, schizophrenia, dissociative identity disorder, depersonalization/derealization disorder, and lucid dreaming), as well as its functional labeling account of the Qualia problem, demonstrate the theory's powerful explanatory power.

This framework not only theoretically resolves the core conflicts between IIT, GWT, and FEP, but more importantly, it transforms the mystery of consciousness into a computable, verifiable, and operational scientific problem. IPWT 2.0 provides a new, unified, and first-principles-driven starting point for understanding the human mind, diagnosing neuropsychiatric disorders, and building artificial general intelligence.

9 Acknowledgements

The formation of this theoretical framework benefited from an interdisciplinary project that explored the intersection of theory construction and narrative possibilities. This project aimed to examine the emergent behavior of formalized theories within complex narrative systems. The author would also like to thank 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," *Conscious. Cogn.*, 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," *Front. Psychol.*, 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," *Nat. Rev. Neurosci.*, vol. 17, no. 5, pp. 307–321, May 2016, doi: 10.1038/nrn.2016.22.

[7] J. Kim, "Making Sense of Emergence ?," *Philos. Stud.*, 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 Comput. Biol.*, vol. 15, no. 2, p. e1006807, 2019, doi: 10.1371/journal.pcbi.1006807.

[9] A. K. Seth and T. Bayne, "Theories of Consciousness," *Nat. Rev. Neurosci.*, 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 Lang.*, 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," *Proc. Natl. Acad. Sci.*, 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," *Nat. Neurosci.*, 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 Comput.*, 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 Neurosci.*, 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?," *Nat. Rev. Neurosci.*, 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," *Sci. Transl. Med.*, 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] A. I. Luppi *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 Comput. Biol.*, 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," *Nat. Rev. Neurosci.*, 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," *Front. Appl. Math. Stat.*, 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 Comput. Biol.*, 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," *Nat. Neurosci.*, 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," *Clin. EEG Neurosci.*, 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," *Neurosci. Conscious.*, 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 Comput. Biol.*, 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?," *Rev. Philos. Psychol.*, 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," *Nat. Neurosci.*, 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?," *Nat. Neurosci.*, 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 Cognit. Sci.*, 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," *Front. Psychol.*, 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 Trans. Neural Netw. Learn. Syst.*, 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," *Conscious. Cogn.*, 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 Trans. Cogn. Develop. Syst.*, 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," *Front. Comput. Neurosci.*, 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," *Nat. Hum. Behav.*, 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," *Front. Syst. Neurosci.*, 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," *Front. Psychol.*, 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 Cognit. Sci.*, 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," *Proc. Natl. Acad. Sci.*, 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?," *J. Exp. Psychol. Gen.*, 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," *Behav. Brain Sci.*, vol. 36, no. 3, pp. 181–204, Jun. 2013, doi: 10.1017/S0140525X12000477.

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

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

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

[65] K. Friston, "A Theory of Cortical Responses," *Philos. Trans. R. Soc. B: Biol. Sci.*, 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," *Biol. Cybern.*, 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 Struct. Funct.*, 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," *J. Hum. Cogn.*, 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 Cognit. Sci.*, 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," *Nat. Rev. Neurosci.*, 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," *Soc. Cogn. Affect. Neurosci.*, 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," *Front. Psychol.*, vol. 9, p. 2714–2715, Jan. 2019, doi: 10.3389/fpsyg.2018.02714.

[75] K. Friston, "Life as We Know It," *J. R. Soc. 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," *Clin. Neurophysiol.*, 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," *Biol. Psychol.*, 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 Lang.*, 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," *Front. 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," *Front. Comput. Neurosci.*, vol. 16, p. 642397–642398, Nov. 2022, doi: 10.3389/fncom.2022.642397.

[84] M. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein, "The Markov blankets of life: autonomy, active inference and the free energy principle," *Journal of The Royal Society Interface*, vol. 15, no. 138, p. 20170792–20170793, Jan. 2018, doi: 10.1098/rsif.2017.0792.

[85] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, 2007. doi: 10.7551/mitpress/4643.001.0001.

[86] J. Hohwy, "The Predictive Mind," 2013, doi: 10.1093/acprof:oso/9780199682737.001.0001.

[87] 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.

[88] 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.

[89] 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

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

[91] 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.

[92] C. M. A. Pennartz, *The Brain's Representational Power: On Consciousness and the Integration of Modalities*. MIT Press, 2015.

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

[94] R. Laukkonen, K. Friston, and S. Chandaria, "A Beautiful Loop: An Active Inference Theory of Consciousness," *Neurosci. Biobehav. Rev.*, p. 106296–106297, Jul. 2025, doi: 10.1016/j.neubiorev.2025.106296.

[95] M. E. Raichle, "The Brain's Default Mode Network," *Annu. Rev. Neurosci.*, vol. 38, no. 1, pp. 433–447, Mar. 2015, doi: 10.1146/annurev-neuro-071714-034853.

[96] A. I. Luppi *et al.*, "Contributions of Network Structure, Chemoarchitecture and Diagnostic Categories to Transitions between Cognitive Topographies," *Nat. Biomed. Eng.*, vol. 8, no. 9, pp. 1142–1161, Aug. 2024, doi: 10.1038/s41551-024-01242-2.

[97] A. I. Luppi *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.

[98] 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.

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

[100] 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 Comput. Biol.*, vol. 17, no. 7, p. e1009196, Jul. 2021, doi: 10.1371/journal.pcbi.1009196.

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

[102] 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

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

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

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

[106] 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 Comput. Biol.*, vol. 20, no. 6, p. e1012178, Jun. 2024, doi: 10.1371/journal.pcbi.1012178.

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

[108] A. I. Luppi, F. E. Rosas, P. A. Mediano, D. K. Menon, and E. A. Stamatakis, "Information Decomposition and the Informational Architecture of the Brain," *Trends Cognit. Sci.*, vol. 28, no. 4, pp. 352–368, Apr. 2024, doi: 10.1016/j.tics.2023.11.005.

[109] J. Rissanen, "Modeling by Shortest Data Description," *Automatica J. IFAC*, vol. 14, no. 5, pp. 465–471, Sep. 1978, doi: 10.1016/0005-1098(78)90005-5.

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

[111] 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

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

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

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

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

[116] 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.

[117] S. Dehaene, C. Sergent, and J.-P. Changeux, "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data during Conscious Perception," *Proc. Natl. Acad. Sci.*, vol. 100, no. 14, pp. 8520–8525, Jul. 2003, doi: 10.1073/pnas.1332574100.

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

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

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

[121] 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

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

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

[124] 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

[125] "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.

[126] 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.

[127] 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.

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

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

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

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

[132] R. L. Carhart-Harris *et al.*, "Neural Correlates of the LSD Experience Revealed by Multimodal Neuroimaging," *Proc. Natl. Acad. Sci.*, vol. 113, no. 17, pp. 4853–4858, Apr. 2016, doi: 10.1073/pnas.1518377113.

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

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

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

[136] 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.

[137] 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.

[138] 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.

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

[140] 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.

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

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

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

[144] 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: Clin.*, vol. 3, pp. 54–64, 2013, doi: 10.1016/j.nicl.2013.07.002.

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

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

[147] E. Selinger, "Reality+: Virtual Worlds and the Problems of Philosophy," *Philos. Mag.*, no. 98, pp. 110–113, 2022, doi: 10.5840/tpm20229875.

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

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

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

[151] C. Schnakers *et al.*, "Cognitive Function in the Locked-in Syndrome," *J. Neurol.*, vol. 255, no. 3, pp. 323–330, Mar. 2008, doi: 10.1007/s00415-008-0544-0.

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

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

[154] 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

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

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

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

[158] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A Survey on Neural Network Interpretability," *IEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 5, pp. 726–742, Oct. 2021, doi: 10.1109/TETCI.2021.3100641.

[159] 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.

[160] 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

[161] 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

[162] 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.

[163] 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.