The Four-Model Theory of Consciousness: A Simulation-Based Framework Unifying the Hard Problem, Binding, and Altered States

Matthias Gruber
Independent researcher
ORCID: 0009-0005-9697-1665
matthias@matthiasgruber.com

Abstract

The science of consciousness remains in a pre-paradigm state, with no theory simultaneously satisfying the eight core requirements a complete theory must meet: the Hard Problem, the Explanatory Gap, the Boundary Problem, the Structure of Experience, Unity and Binding, Combination and Emergence, the Causal Role, and the Meta-Problem. This paper presents the Four-Model Theory, in which consciousness is constituted by real-time self-simulation across four nested models arranged along two axes—scope (world vs. self) and mode (implicit vs. explicit). The implicit models (Implicit World Model, Implicit Self Model) are substrate-level, learned, and non-conscious. The explicit models (Explicit World Model, Explicit Self Model) are virtual, transient, and phenomenal—they are the simulation in which experience occurs. The theory's central claim is that qualia are virtual: they are the way the simulated self perceives its own states within the simulation, not properties of the physical substrate. This dissolves the Hard Problem by showing that "Why does physical processing feel like something?" commits a category error—the physical processing does not feel; the simulation does, and within the simulation, qualia are constitutive. Combined with a criticality requirement (the substrate must operate at the edge of chaos), the theory derives diverse phenomena from five principles: criticality, virtual qualia, a redirectable Explicit Self Model, variable implicit–explicit permeability, and virtual model forking. These principles unify psychedelic phenomenology, anesthetic mechanisms, dream states, split-brain phenomena, dissociative identity disorder, and animal consciousness. A systematic comparison shows the theory addresses all eight requirements. Nine novel testable predictions are offered, including that psychedelic ego dissolution content is controllable via sensory input and that psychedelics should alleviate anosognosia—predictions no competing theory generates. The criticality requirement, derived from Wolfram's computational framework in 2015 independently of (though not prior to) the

empirical criticality program, converges with empirical criticality literature consolidated in 2025–2026 (Hengen and Shew, 2025; Algom and Shriki, 2026)—a convergence from theoretical and empirical starting points.

Keywords: consciousness, hard problem, self-model, simulation, qualia, criticality, binding problem, altered states, psychedelics, substrate independence

1 Introduction

1.1 The Pre-Paradigm State of Consciousness Science

After three decades of intensive scientific investigation, consciousness research finds itself at an impasse. The field possesses no dominant paradigm in the Kuhnian sense (Kuhn, 1962), no agreed-upon methodology for linking subjective experience to objective measurement, and no theory that commands broad assent. Multiple frameworks compete for explanatory primacy—Integrated Information Theory (IIT; Tononi, 2004; Albantakis et al., 2023), Global Neuronal Workspace (GNW; Baars, 1988; Dehaene and Changeux, 2011), Higher-Order Theories (HOT; Rosenthal, 2005; Lau and Rosenthal, 2011), Predictive Processing (PP; Friston, 2010; Seth, 2021), Attention Schema Theory (AST; Graziano, 2013), Recurrent Processing Theory (RPT; Lamme, 2006, 2010), and others—yet none has established decisive empirical or theoretical superiority over its rivals.

Recent developments have deepened the crisis rather than resolving it. The COGITATE adversarial collaboration—whose protocol was pre-registered by Melloni et al. (2023) and whose results were published in *Nature* by the COGITATE Consortium (2025)—produced equivocal results: neither IIT nor GNW was fully confirmed, and the data favored posterior cortical involvement over either camp's predictions. A letter signed by over 100 researchers declared IIT pseudoscientific (IIT-Concerned et al., 2025; Nature Neuroscience Editors, 2025), provoking fierce rebuttals (Tononi et al., 2025) and calls for more nuanced framing (Gomez-Marin and Seth, 2025). Reviews published in 2024–2025 have asked whether the field is making genuine progress or merely accumulating incompatible frameworks.

This paper argues that the impasse persists because no existing theory simultaneously addresses all of the fundamental requirements that a complete theory of consciousness must meet. Each theory excels on some requirements but remains silent on, or weak against, others. The field needs a framework that addresses the full set of challenges—not just neural correlates or access conditions, but the deep philosophical problems that make consciousness uniquely difficult.

1.2 What Would Count as Progress?

I propose that any theory claiming to provide a comprehensive account of consciousness must address eight core requirements, drawn from the philosophical and scientific literature. These requirements are not novel; each has been identified by previous authors as a central challenge. What is novel is the demand that a single theory address all eight simultaneously:

1. **The Hard Problem** (Chalmers, 1995)—Why does physical processing give rise to subjective experience?
2. **The Explanatory Gap** (Levine, 1983)—Why does the explanation of neural correlates feel incomplete?
3. **The Boundary Problem** (Bayne, 2010; Tononi, 2004)—Where does the conscious system end?
4. **The Structure of Experience** (Nagel, 1974)—How does physical processing produce richly structured experience?
5. **Unity and Binding** (Treisman and Gelade, 1980; Revonsuo, 1999)—How are distributed processes unified into coherent experience?
6. **Combination and Emergence** (Chalmers, 2016)—How do non-conscious elements combine to produce consciousness?
7. **The Causal Role** (Jackson, 1982)—Does consciousness do anything?
8. **The Meta-Problem** (Chalmers, 2018)—Why do we think there is a hard problem?

Section 2 develops each requirement in detail and surveys how existing theories fare against them. The remainder of the paper presents a theory—the Four-Model Theory—that, I argue, addresses all eight.

1.3 Overview and Historical Context

The Four-Model Theory was originally published in German as *Die Emergenz des Bewusstseins* (Gruber, 2015) and has been refined through a structured adversarial challenge process in 2026. The theory self-identifies as an intersection of Dennett's Multiple Drafts Model (Dennett, 1991), Metzinger's Self-Model Theory of Subjectivity (Metzinger, 2003, 2009), and neural network architecture. It is substrate-independent: the six-layer mammalian cortex is understood as an evolutionary implementation, not a requirement.

The theory proposes that consciousness consists of a real-time self-simulation running on an

implicit (substrate-level) knowledge base. Qualia—the subjective "feel" of experience—are virtual: they exist within and are constitutive of the simulation, but do not exist at the substrate level. This two-level ontology dissolves the Hard Problem by showing it rests on a category error.

Combined with a criticality requirement derived independently from Wolfram's computational framework (Wolfram, 2002), the theory generates nine novel testable predictions and unifies phenomena across psychopharmacology, clinical neurology, sleep science, and comparative cognition under five principles.

The theory is offered as one model among several, contributing to humanity's collective search for an adequate account of consciousness. Every model carries inherent modeling error; the present theory is no exception. It is intended to complement existing frameworks—extending where they are incomplete, not displacing where they succeed.

The paper proceeds as follows. Section 2 establishes the eight requirements. Section 3 presents the Four-Model Theory. Sections 4 and 5 develop the philosophical commitments and the binding/criticality framework. Section 6 demonstrates the theory's explanatory range. Section 7 provides a systematic comparative analysis against major competitors. Section 8 presents the nine predictions. Sections 9–11 address open questions, implications, and conclusions.

2 Eight Requirements for a Theory of Consciousness

This section develops each of the eight requirements and briefly notes how current theories address or fail to address them. A detailed theory-by-theory comparison follows in Section 7; the purpose here is to establish the evaluative framework.

2.1 The Hard Problem

Chalmers (1995, 1996) formulated the Hard Problem as the question of why physical processing is accompanied by subjective experience. We can explain all the *functions* of consciousness—discrimination, integration, reporting, attention—without explaining why there is "something it is like" (Nagel, 1974) to undergo these processes. The explanatory challenge is not about identifying neural correlates but about understanding why correlates are accompanied by phenomenality at all.

Most neuroscientific theories of consciousness (GNW, RPT, PP) focus on functional or mechanistic accounts and remain explicitly or implicitly silent on the Hard Problem. IIT

attempts to address it by defining consciousness as intrinsic causal power (Φ), treating experience as identical to integrated information—but this requires accepting panpsychist commitments that many find problematic (Aaronson, 2014; Doerig et al., 2019). HOT and AST offer deflationary accounts that explain *why we report* having phenomenal experience but leave open whether they have truly addressed the phenomenality itself. Illusionism (Dennett, 1991; Frankish, 2016) dissolves the problem by denying that qualia as traditionally conceived exist—a position that remains deeply controversial.

2.2 The Explanatory Gap

Levine (1983) identified the Explanatory Gap as a distinct problem from the Hard Problem: even if we identify every neural correlate of every conscious state, the explanation seems to leave something out. The gap is between the third-person description (neural firing patterns) and the first-person reality (what the experience is like). Block (1995, 2007) further refined this as the distinction between access consciousness (the functional role) and phenomenal consciousness (the subjective feel).

The Explanatory Gap is often treated as a restatement of the Hard Problem, but it has a distinct character: it is about the *form* of explanation rather than the *existence* of the phenomenon. A theory that dissolves the Hard Problem should simultaneously close the Explanatory Gap.

2.3 The Boundary Problem

Where does the conscious system begin and end? Within the brain, only some processing is conscious at any given moment. Between organisms, it is unclear where to draw the line. The Boundary Problem asks for a principled account of what delineates conscious from non-conscious processing (Bayne, 2010; Tononi, 2004).

IIT provides the strongest existing treatment of this requirement through its exclusion postulate: the system with maximum Φ defines the boundary of consciousness. However, the computational intractability of calculating Φ limits its practical application. GNW defines conscious access in terms of global broadcasting, but the boundary between broadcast and non-broadcast content is not always sharp. PP uses Markov blankets (Friston, 2010) but has been criticized for being too liberal in its boundary-setting (Bruineberg et al., 2022).

2.4 The Structure of Experience

Conscious experience is not a homogeneous blob—it has rich spatial, temporal, modal, and qualitative structure. A visual scene has colors, shapes, depths, and textures; an auditory experience has pitches, timbres, and spatial locations; emotional experience has valence, intensity, and phenomenal character. Any complete theory must explain how physical processing generates this structured phenomenology.

IIT's qualia space provides a mathematical treatment of experiential structure, arguably its greatest strength. PP's generative models are inherently structured, providing a natural account of the richness of perceptual experience. GNW and HOT are weaker here, offering accounts of *when* content becomes conscious but less about *why* it has the particular structure it does.

2.5 Unity and Binding

The Binding Problem (Treisman, 1996; Revonsuo, 1999) asks how distributed neural processes—occurring in different brain regions, at different timescales, in different modalities—are unified into a single coherent conscious experience. I see a red ball: "red," "round," "ball," "there," and "now" are processed in different cortical areas, yet I experience a unified percept.

Proposed solutions range from temporal synchrony (Gray et al., 1989; Singer and Gray, 1995; Fries, 2005, 2015) to integrated information (Tononi, 2004) to global broadcasting (Baars, 1988). None is universally accepted. The binding problem remains, alongside the Hard Problem, one of the deepest unresolved challenges.

2.6 Combination and Emergence

How do non-conscious elements combine to produce consciousness? For panpsychist theories (IIT in its strong form; Goff, 2019; Strawson, 2006), this takes the form of the Combination Problem (James, 1890; Chalmers, 2016): if fundamental entities have micro-experience, how do these micro-experiences combine into the macro-experience we know? For physicalist theories, the challenge is one of emergence: at what point, and by what mechanism, does consciousness emerge from non-conscious physical processes?

The Combination Problem is widely regarded as panpsychism's most serious difficulty (Chalmers, 2016; Coleman, 2014). Physicalist emergence theories face the objection that they either invoke strong emergence (which many philosophers consider mysterious) or reduce consciousness to function (which many consider inadequate). A satisfactory theory must

navigate between these difficulties.

2.7 The Causal Role of Consciousness

Does consciousness *do* anything, or is it epiphenomenal—a by-product with no causal power? If consciousness has causal power, what kind? If it does not, why does it exist?

This requirement is politically charged within the field. Epiphenomenalism (Huxley, 1874; Jackson, 1982) is widely dismissed as absurd (how could evolution produce something causally inert?), yet many mechanistic theories implicitly struggle with the question. If consciousness is "just" a global broadcast (GNW) or "just" recurrent processing (RPT), what role does the *experience* play beyond the mechanism? The PP framework, through active inference, provides perhaps the strongest existing case for consciousness having a functional role, but it is unclear whether the functional role requires phenomenal consciousness rather than mere information processing.

2.8 The Meta-Problem

Chalmers (2018) identified the Meta-Problem: why do we *think* there is a Hard Problem? Even if the Hard Problem is illusory (as illusionists argue) or misformulated (as functionalists hold), it is a fact that most humans—including most philosophers and scientists—report a strong intuition that consciousness is deeply mysterious. A complete theory should explain this intuition.

AST provides the strongest existing account of the Meta-Problem: we model our own attention, and the model necessarily omits the mechanistic details, leading to the intuition that consciousness is something non-physical (Graziano, 2013). This is a genuine insight. However, AST's treatment of the Meta-Problem does not extend to a solution of the Hard Problem itself—it explains why we *think* there is a mystery without fully accounting for the mystery.

3 The Four-Model Theory

3.1 Core Definition

Consciousness is the ability of an entity—biological or artificial—to create a model of itself, to relate that model to itself, and to interact with it. Consciousness is not a property the brain possesses but a process the brain performs: it runs a real-time self-simulation.

This definition is functional and substrate-independent. It does not require a specific physical implementation, biological composition, or computational architecture. What it requires is a system capable of constructing and maintaining a self-referential simulation in real time.

3.2 The Four Models

The theory identifies four nested models distinguished by two orthogonal dimensions: **scope** (everything vs. self only) and **mode** (implicit/learned vs. explicit/simulated). This is a conceptual taxonomy, not a claim about spatial organization in the brain—the models are functionally distinct processes, not anatomically localized regions.

Table 1: The Four-Model Architecture

	Everything (world)	**Self only**
Implicit (learned, substrate-level)	Implicit World Model (IWM)	Implicit Self Model (ISM)
Explicit (simulated, phenomenal)	Explicit World Model (EWM)	Explicit Self Model (ESM)

The Implicit World Model (IWM) encompasses the substrate's total accumulated knowledge about the world, stored in synaptic weights (or their functional equivalent in non-biological substrates). It includes everything the system has ever learned: perceptual regularities, causal models, spatial relationships, semantic knowledge, motor programs for interacting with the world. The IWM is never directly conscious. It operates "in the dark"—providing the knowledge base from which the conscious simulation is generated, but never itself appearing in experience.

The Implicit Self Model (ISM) is the substrate's accumulated self-knowledge: body schema, proprioceptive calibration, motor skills, habits, personality traits, autobiographical memory structures, and social self-knowledge. Like the IWM, the ISM is never directly conscious. There is no unified homunculus—no inner observer reading the ISM. The ISM is a structural feature of the substrate, not an experiential one.

The Explicit World Model (EWM) is the conscious world—the real-time simulation of reality that constitutes perceptual experience. When you see a room, hear a voice, feel the texture of a surface, you are experiencing the EWM. It is generated from the IWM (which provides the world-knowledge) and current sensory input (which constrains and updates the simulation), but it is not identical to either. The EWM is a virtual construct—a transient pattern of activity, not a permanent structure.

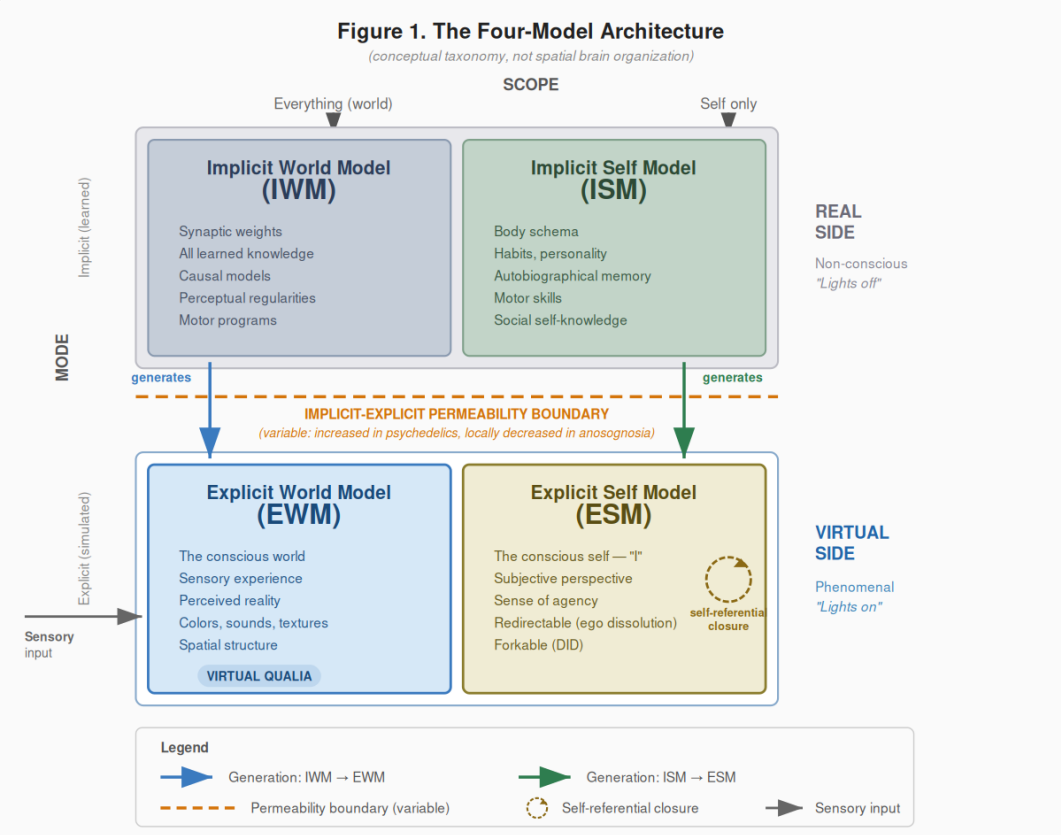

Figure 1: The four-model architecture. The two orthogonal axes—scope (world vs. self) and mode (implicit vs. explicit)—define four functionally distinct models. Implicit models (bottom) are substrate-level, learned, and non-conscious. Explicit models (top) are virtual, transient, and phenomenal.

The Explicit Self Model (ESM) is the conscious self—the real-time simulation of "I" that constitutes self-experience. It is the sense of being a subject, having a perspective, occupying a body, possessing a history, and being the author of one's actions. The ESM is generated from the ISM (which provides the self-knowledge) and current interoceptive and proprioceptive input, but like the EWM, it is virtual: a transient process, not a permanent entity.

3.3 The Real/Virtual Split

The four models divide into two fundamental categories:

The real side (IWM + ISM): These are physical, structural, learned, and non-conscious. They are stored in the substrate's architecture—in biological brains, primarily in synaptic weights, dendritic morphology, and connectivity patterns. They accumulate over the organism's lifetime through learning. They have no phenomenal character. "Lights off."

The virtual side (EWM + ESM): These are simulated, transient, generated, and phenomenal. They are patterns of activity—in biological brains, transient electrochemical dynamics. They are constructed in real time from the implicit models and current sensory input. They *are* experience. "Lights on."

This division is the foundation of the theory's treatment of the Hard Problem (Section 3.4) and structures its account of every phenomenon it addresses.

The virtual models possess **software-like properties** that follow from their nature as simulations rather than structures:

- **They can be forked**: A single substrate can run multiple configurations of the ESM (see Section 6.2 on dissociative identity disorder).
- **They can be cloned**: Physical separation of the substrate produces degraded but complete copies of the virtual models (see Section 6.4 on split-brain).
- **They can be redirected**: The ESM requires input; disrupt normal self-referential input and it latches onto whatever input dominates (see Section 6.1 on psychedelics).
- **They can be reconfigured**: Therapeutic interventions (CBT, exposure therapy) work by modifying the virtual models through substrate-level rewiring (see Section 6.6).

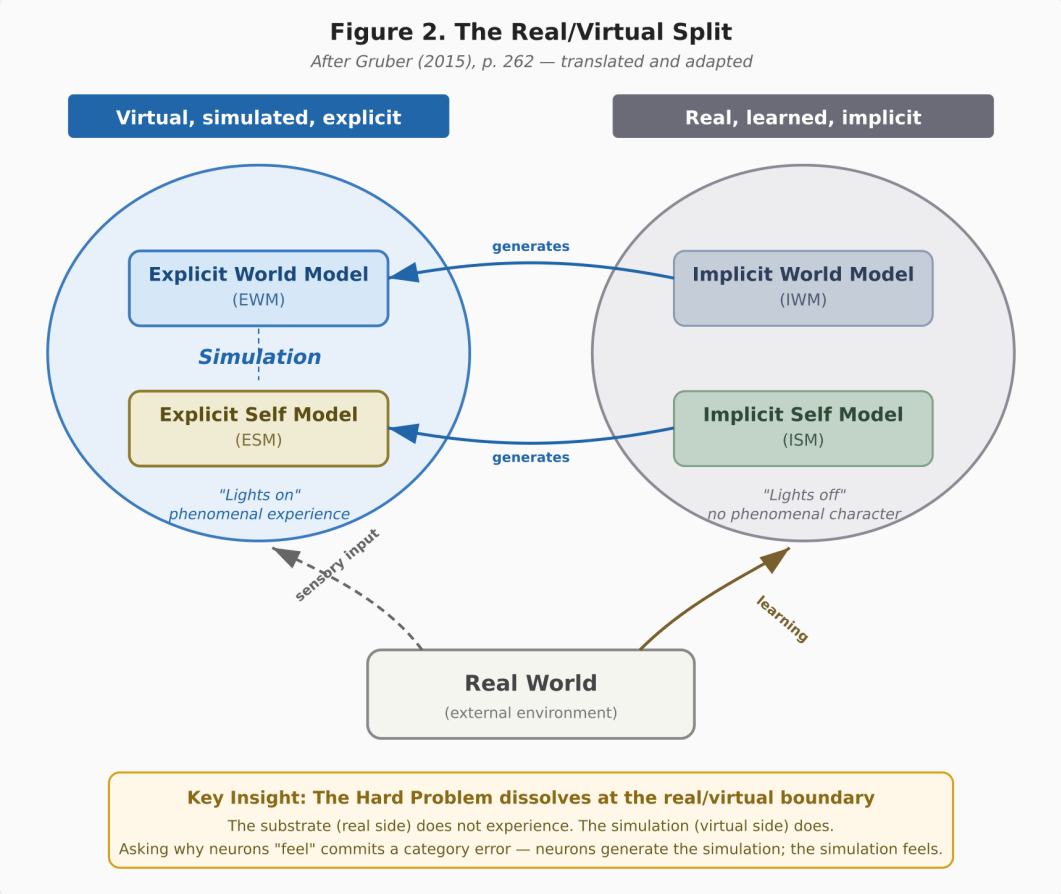

Figure 2: The ontological split between the real substrate (physical, structural, non-conscious—"lights off") and the virtual phenomenal world (simulated, transient, experiential—"lights on"). Qualia exist only on the virtual side.

3.4 Virtual Qualia: Dissolving the Hard Problem

The central claim of the Four-Model Theory is that **qualia are virtual**. They are the way the simulated self (ESM) perceives its own states and the simulated world (EWM). Qualia exist within and are constitutive of the simulation; they do not exist at the substrate level.

This dissolves the Hard Problem by revealing a category error in its formulation:

The standard formulation: "Why does physical processing (neuronal firing, synaptic transmission) feel like something?"

The dissolution: The physical processing *does not* feel like anything. The IWM and ISM—the substrate-level implicit models—operate without any phenomenal character whatsoever. There is nothing it is like to be a synaptic weight. The simulation, however, *does* feel—and within the simulation, qualia are simply what self-perception produces. Asking why neuronal firing feels like something is analogous to asking why transistor switching feels like running a video game. The transistors do not run the game at the level of individual switching; the virtual machine does. The neurons do not experience redness at the level of individual firing; the simulation does, and within the simulation, "redness" is simply the ESM's mode of registering a particular class of EWM content.

Why self-simulation specifically? A critic might object that this merely relocates the Hard Problem: why does *this* virtual process have experience when a weather simulation does not? The answer lies in **self-referential closure**. A weather simulation models weather; it does not model *itself modeling weather*. The four-model architecture creates a closed loop: the ESM is the system modeling its own modeling process. In this loop, the distinction between the model and the modeled collapses—the simulation *is* the thing being simulated. Qualia are not an *addition* to the self-modeling; they are the self-modeling as encountered from the inside of the loop. A non-self-referential simulation has an outside from which it can be described without remainder; a self-referential simulation at criticality has no such outside. The simulation *is* its own observer, and observation-from-inside is what we call experience.

This is not a proof that self-referential simulation must be conscious—it is an argument that self-referential simulation is the *kind* of process for which the Hard Problem's assumptions break down. Self-referential closure is precisely the condition under which the gap between process and feeling does not exist.

This is **not** illusionism in the sense of Dennett (1991) or Frankish (2016) or Graziano (2024). Illusionism holds that qualia as traditionally conceived are illusions—there is nothing it is like, and our sense that there is something it is like is itself a misrepresentation. The Four-Model

Theory holds that qualia are *real within the simulation*. Within the EWM/ESM, experience has genuine phenomenal character. What is illusory is the assumption that this phenomenal character must be a property of the physical substrate. It is not. It is a property of the virtual process that the substrate runs.

This constitutes a **two-level ontology**: the substrate level (real side) has no experience, and the simulation level (virtual side) has genuine experience. Both levels are physical—the simulation is a physical process, not a supernatural one—but they have different ontological properties. The category error in the Hard Problem consists in conflating the two levels: seeking phenomenal properties at the substrate level where they do not exist.

The Explanatory Gap closes simultaneously. The gap between "neurons fire in pattern X" and "I experience red" is not a gap in our knowledge but a reflection of the level distinction. The neural firing pattern generates and sustains the simulation in which redness is experienced, but the firing pattern itself is not red and does not experience redness, just as a CPU's electrical states are not "a spreadsheet" even though they generate and sustain one.

3.5 Graduated Levels of Consciousness

Consciousness in the Four-Model Theory is not binary but graduated. The theory identifies a hierarchy of levels based on the depth of recursive self-modeling:

Basic consciousness: Minimal self-simulation. The system generates an EWM and a rudimentary ESM—it experiences a world and has a minimal sense of being a subject within it. This is the entry level: phenomenal experience exists but self-awareness is thin.

Simply extended consciousness: First-order self-observation. The system models itself—the ESM includes a model of the system's own states and processes. The organism not only experiences but is aware that it experiences.

Doubly extended consciousness: Second-order self-observation. The system models itself modeling itself. This enables reflection, metacognition, and the sense of being an observer of one's own mental processes.

Triply extended consciousness: Third-order self-observation. The system models itself modeling itself modeling itself. This supports the deepest forms of self-awareness, philosophical reflection, and the very intuition that consciousness is mysterious (connecting to the Meta-Problem—see Section 3.8). Notably, triply extended consciousness is also a prerequisite for the scientific study of consciousness itself: only a system capable of modeling its own modeling of its own experience can formulate the question "What is consciousness?"

Each level corresponds to an additional layer of recursive self-modeling. The levels are not discrete stages but points along a continuum. Different organisms—and potentially different artificial systems—occupy different positions along this continuum, and individual organisms may fluctuate between levels depending on state (waking, dreaming, meditative, intoxicated).

3.6 The Implicit–Explicit Boundary

A key mechanism in the Four-Model Theory is the **permeability of the boundary between implicit models (IWM/ISM) and explicit models (EWM/ESM)**. Information becomes conscious when it is transferred from the implicit to the explicit side—when substrate-level knowledge or self-knowledge is incorporated into the running simulation.

In normal waking states, this boundary is **selectively permeable**: relevant information passes through based on attentional and contextual gating. You are not conscious of everything your IWM knows about the world or everything your ISM knows about yourself; you are conscious of what the current simulation requires.

The permeability of this boundary is variable, and its variation explains a wide range of phenomena (detailed in Section 6):

- **Psychedelic states**: Global increase in permeability—intermediate processing stages (normally implicit) become accessible to the explicit models. This explains the characteristic visual progression from simple phosphenes (V1-level) through geometric patterns (V2/V3-level) to complex imagery (higher visual areas) and eventually full dream-like scenes (Carhart-Harris et al., 2014).
- **Anosognosia**: Local decrease in permeability—the ISM contains the information (e.g., that a limb is paralyzed), but the transfer to the EWM is blocked for that specific domain.
- **Pre-sleep/deep relaxation**: Gradually increasing permeability, producing the same bottom-up visual progression as psychedelics (phosphenes → geometrics → hypnagogic imagery).
- **Meditation**: Trained modulation of permeability, enabling selective access to normally implicit processes.

Importantly, the implicit–explicit boundary is not a sharp line but a **graded transition zone**. Behavioral complexity itself follows a gradient—from reflexive chemical-gradient responses through conditioned, goal-directed, and rule-based behavior to fully conscious action—and the implicit and explicit memory systems overlap precisely in the middle of this gradient,

at the levels of goal-directed and template-based behavior (Gruber, 2015). This overlap zone is the functional locus of the variable permeability described above: behavior at these intermediate levels can be driven by either implicit or explicit processing, depending on attentional state, arousal, and contextual demands.

3.7 The Criticality Requirement

The Four-Model Theory imposes a **physical prerequisite** for consciousness: the substrate must operate at or near the edge of chaos—Wolfram's Class 4 computational regime (Wolfram, 2002).

Wolfram classified cellular automata (and by extension computational systems generally) into four classes:

- **Class 1**: Converges to a fixed state. Too simple for consciousness.
- **Class 2**: Periodic/repetitive. Too simple for consciousness.
- **Class 3**: Chaotic/random. Too disordered for coherent consciousness.
- **Class 4**: Complex/edge of chaos. Capable of universal computation. The regime in which consciousness can emerge.

This classification was applied to the question of consciousness in Gruber (2015), where it was argued that consciousness requires Class 4 dynamics—complex enough to sustain a self-simulation, ordered enough for that simulation to be coherent. This requirement was derived *theoretically*, from the computational properties needed for real-time self-modeling, not from empirical neuroscience.

Independently, empirical neuroscience has converged on the same conclusion through a different path. Beggs and Plenz (2003) demonstrated neuronal avalanches consistent with self-organized criticality in cortical tissue. Carhart-Harris et al. (2014) proposed the Entropic Brain Hypothesis, linking consciousness level to neural entropy. Tagliazucchi et al. (2012, 2016) showed criticality signatures in waking fMRI and under LSD. Priesemann et al. (2013, 2014) characterized brain dynamics as slightly subcritical in normal waking states. This line of research was formally consolidated in the Consciousness and Criticality (ConCrit) framework (Algom and Shriki, 2026), which synthesized evidence across multiple paradigms to establish that consciousness tracks criticality across pharmacological, pathological, and physiological state changes. A complementary meta-analysis of 140 datasets confirmed criticality as a unified setpoint of brain function (Hengen and Shew, 2025).

Table 2: Independent Convergence on Criticality

Year	Development	Path
2002	Wolfram publishes *A New Kind of Science*	Computational theory
2003	Beggs & Plenz—neuronal avalanches	Empirical neuroscience
2014	Carhart-Harris—Entropic Brain Hypothesis	Empirical neuroscience
2015	**Gruber—Class 4 / edge of chaos requirement**	**Theoretical (via Wolfram)**
2016	Tagliazucchi et al.—LSD and criticality	Empirical neuroscience
2022	"Self-organized criticality as a framework" (review)	Empirical neuroscience
2025	Hengen & Shew—meta-analysis of 140 datasets	Empirical neuroscience
2025–26	**ConCrit framework (Algom & Shriki)**	**Theoretical/empirical synthesis**

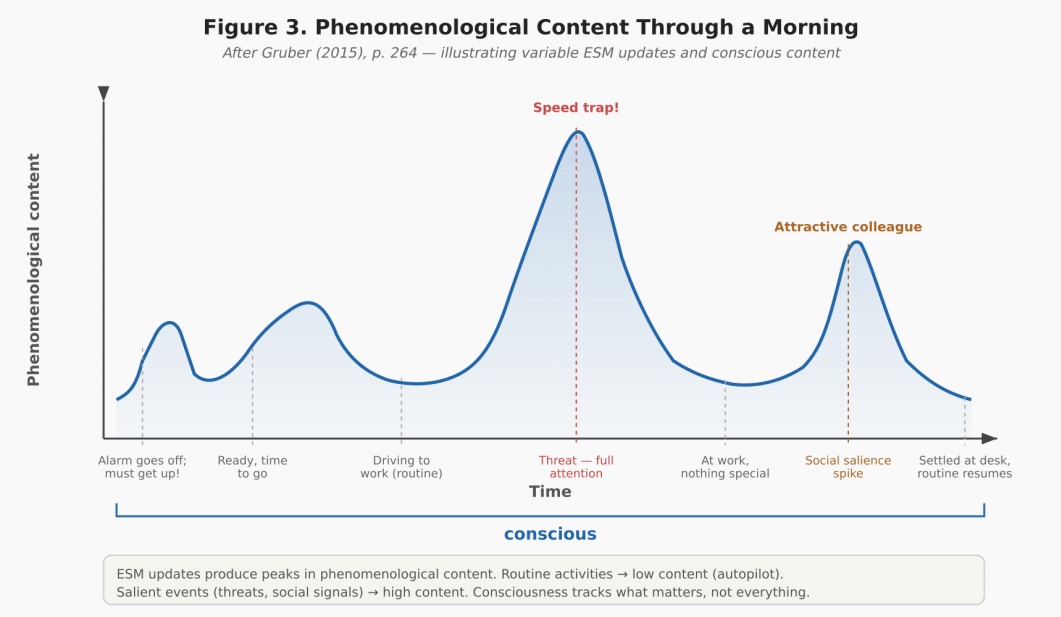

Figure 3: The structure of phenomenological content: what appears in the virtual world (EWM) and how the virtual self (ESM) experiences it. The boundary between implicit and explicit determines what reaches conscious awareness.

Two paths—theoretical reasoning from Wolfram's computational universality framework (Gruber, 2015) and large-scale empirical neuroscience (Hengen and Shew, 2025; Algom and Shriki, 2026)—converging on the same claim. Although the empirical criticality program was already underway (Beggs and Plenz, 2003) when Gruber (2015) derived the requirement from computational first principles rather than from neuroimaging data, this convergence does not prove the Four-Model Theory correct, but it provides notable support for one of its core predictions.

The Four-Model Theory distinguishes two thresholds for consciousness:

- **Physical threshold**: Criticality. The substrate must operate at Class 4 dynamics. Below this, no consciousness is possible regardless of architecture.
- **Functional threshold**: Four-model architecture. The substrate must implement the four-model self-simulation. Above criticality but without the architecture, there is complex dynamics but no consciousness.

Both thresholds must be met. Criticality is necessary but not sufficient; the four-model architecture is necessary but not sufficient. Together they are sufficient.

3.8 The Meta-Problem Dissolved

The Meta-Problem—why we think there is a Hard Problem—receives a natural account within the Four-Model Theory. The ISM (Implicit Self Model) is **structurally inaccessible** to the ESM (Explicit Self Model). The conscious self cannot directly observe its own substrate. When the ESM attempts to model the basis of its own experience, it encounters a principled opacity: the implicit models that generate the simulation are not themselves part of the simulation.

This is why consciousness *seems* mysterious. The ESM can represent that it is having an experience, but it cannot represent the mechanism by which the experience is generated—because that mechanism operates at the implicit/substrate level, which is by definition outside the explicit/virtual level. The result is the persistent intuition that something is being "left out" of any physical explanation: the ESM cannot find the mechanism within its own simulation, so it concludes the mechanism must be non-physical or fundamentally inexplicable.

This account shares features with Graziano's AST explanation—both invoke the self-model's necessary incompleteness—but grounds it in a more specific architecture (four models, real/virtual split) and connects it to the broader dissolution of the Hard Problem rather than

treating the Meta-Problem in isolation.

4 Philosophical Commitments

The Four-Model Theory entails specific philosophical positions that were established through structured adversarial analysis and are internally consistent. This section develops and defends each commitment.

4.1 Process Physicalism

The theory is physicalist: both the substrate (implicit models) and the simulation (explicit models) are physical processes. There is no non-physical substance, no fundamental experiential property, no panpsychist micro-experience. Qualia are higher-order physical patterns—specifically, they are patterns of activity within the simulation that constitute the ESM's self-perception within the EWM.

This is process physicalism rather than identity theory: consciousness is not identical to any particular neural state but is constituted by the *process* of self-simulation. The same conscious state could, in principle, be realized by different physical substrates—what matters is the functional architecture (four models at criticality), not the specific material.

Process physicalism avoids the difficulties of both type-identity theory (which struggles with multiple realization) and functionalism as traditionally conceived (which struggles with the Hard Problem). The Four-Model Theory adds the real/virtual level distinction to standard functionalism, which is what allows it to address phenomenality: qualia are not just functional roles but virtual properties of the simulation. They are real *as virtual properties*—genuinely experiential but not properties of the substrate.

4.2 Consciousness as Process, Not Agent

A persistent source of confusion in consciousness studies is the treatment of consciousness as an entity—an "it" that either does or does not cause things. The Four-Model Theory rejects this framing. Consciousness is not a thing; it is a process *performed* by the substrate. Asking whether consciousness "causes" anything is a category error—analogous to asking whether the pointer of a clock meeting the numerals causes the clock to work. The energy source drives the gears, which drive the pointer, but nowhere does the virtual interaction between pointer and numeral cause anything mechanical. Yet without that interaction the clock cannot be said to function—or malfunction.

The implicit models generate the virtual simulation for concrete adaptive reasons: the EWM integrates multimodal sensory data into a unified scene; the ESM provides a self-model against which consequences can be evaluated. This is not idle accompaniment. An experience that felt aversive updates the implicit models to avoid similar situations; an experience of successful agency reinforces the motor and social patterns that produced it. The virtual simulation is the substrate's mechanism for consequence-observation and future-oriented adaptation—the very thing natural selection shaped the architecture to do.

This makes the theory's position distinct from classical epiphenomenalism, in which consciousness is a causally inert by-product with no functional role. In the Four-Model Theory, the virtual models are in continuous feedback with the implicit models: the simulation's outputs feed back to update implicit processing, shaping future behavior. Qualia, as constitutive elements of that simulation, lack independent causal power over the substrate—much as the hands and numerals of a clock have no direct mechanical relation to the gear train, yet the clock cannot function as a clock without them. Remove the display and the mechanism still runs, but it no longer serves its purpose.[1]

The theory also reframes the free will debate. The ESM narrates decisions already made at the substrate level (Libet, 1985; Schurger et al., 2012; Wegner, 2002), which might seem to eliminate free will—but only if "will" is restricted to what the conscious self explicitly desires. The Four-Model Theory suggests a broader view: the substrate, including the implicit models, continuously optimizes the organism's existence, and this optimization *is* the individual's will—merely not fully transparent to the ESM. One's will is real but only partially known to oneself. The conscious experience of wanting something is the ESM's window onto a deeper process that is genuinely goal-directed. Whether this constitutes free will in the libertarian sense reduces to a question of physical determinism—a question for physics, not consciousness theory. But the theory predicts that even extreme acts of will, including self-destruction, reflect the system's optimization rather than its failure—which is, paradoxically, among the stronger arguments that the will is genuine.

This framing addresses the standard objections. **Zombies** (Chalmers, 1996): not possible—the virtual models *are* the substrate's activity at the virtual level, just as a vortex is not added to water's movement but is a description of it. A system implementing the four-model architecture at criticality necessarily instantiates the simulation, and the simulation necessarily has phenomenal character. **The knowledge argument** (Jackson, 1982): Mary

[1] As a corollary, consciousness plays no role in quantum measurement or wavefunction collapse. Observer-dependent interpretations of quantum mechanics (von Neumann, 1932; Wigner, 1961) are rejected. This is consistent with the dominant position in contemporary physics (decoherence approaches; Zurek 2003) and with the theory's commitment to physicalism.

gains acquaintance with a virtual quale she could not access from substrate descriptions—real knowledge, no independent causal power required. **The evolutionary argument**: natural selection targets the functional capabilities of the architecture (predictive modeling, social cognition, consequence-evaluation); the phenomenal character of the simulation is constitutive of those capabilities, not a separate target and not a free-rider.

4.3 Weak Emergence

Consciousness, in this theory, is weakly emergent: it is deducible in principle from a complete description of the substrate, even if it is practically irreducible due to the complexity of the system. There is no strong emergence, no magical threshold, no point at which "something extra" appears that could not have been predicted from the underlying physics.

This avoids the difficulties of both strong emergence (which is either mysterious or incoherent; Kim, 1993) and the panpsychist combination problem (which is unresolved; Chalmers, 2016). Consciousness does not arise from combining micro-experiences (there are none to combine) and does not require a special emergence law (there is none). It arises from the computational properties of a system running a self-simulation at criticality, just as a weather pattern arises from the thermodynamic properties of an atmosphere—no extra ingredient needed.

4.4 Substrate Independence

The six-layer mammalian neocortex is an evolutionary implementation of the four-model architecture, not a requirement for it. Consciousness is substrate-independent: any physical system capable of implementing the four-model architecture at criticality should produce consciousness.

Biological evidence already supports this. Corvids (crows, ravens) and parrots demonstrate cognitive abilities—tool use, planning, mirror self-recognition, social cognition—that strongly suggest consciousness, yet their brains have no neocortex. Their pallium is organized in nuclear clusters rather than layers (Güntürkün and Bugnyar, 2016). Cephalopods (octopuses) demonstrate problem-solving and behavioral flexibility with an even more radically different brain architecture. If the Four-Model Theory is correct, these animals are conscious not because they share our neural architecture but because they have evolved functionally equivalent self-simulation architectures on different substrates—exactly what substrate independence predicts.

The implication for artificial consciousness is direct: a synthetic system implementing the four-model architecture at criticality should produce genuine consciousness. Current AI systems,

including large language models, do not meet this specification. LLMs lack an Explicit Self Model (they do not run a real-time self-simulation), lack criticality (transformer inference is a feedforward pass—Class 1/2 dynamics), and lack the real/virtual split that grounds phenomenality. The theory predicts that the qualitative difference between interacting with a genuinely conscious artificial system and interacting with an LLM would be immediately and qualitatively distinguishable.

5 Binding, Criticality, and Holographic Storage

5.1 Binding as an Emergent Property of Critical Dynamics

The Four-Model Theory does not treat binding as a separate mechanism requiring a dedicated solution. Instead, binding is an emergent property of a substrate operating at criticality.

At the edge of chaos, the system exhibits maximal correlation length: distant parts of the substrate influence each other, local perturbations propagate globally, and information is integrated across the entire network. These are precisely the conditions under which distributed representations—features encoded across the full network rather than localized in specific neurons—are sustained and coordinated. Binding is not something the brain does *in addition to* its other computations; it is a consequence of the dynamical regime in which the brain operates.

This is consistent with empirical findings. Gamma-band synchrony (Gray et al., 1989; Rodriguez et al., 1999; Fries, 2005, 2015), long-range thalamocortical coherence (Llinás and Ribary, 1993; Llinás et al., 1998), and criticality signatures in neural dynamics (Beggs and Plenz, 2003; Tagliazucchi et al., 2012) all point toward the same conclusion: the unified character of conscious experience reflects the dynamical integration of a system operating at or near criticality, not a dedicated binding mechanism.

A further role for criticality emerges from the recursive structure of the four-model architecture. The theory requires that models of different orders of complexity—from a basic world model (EWM) to a self-model observing a self-model (ESM monitoring ESM)—coexist and interact in real time. This demands cross-scale synchronization: simple and complex representations must remain coherent with one another. Criticality's scale-free dynamics provide exactly this, supporting information flow across all levels of the representational hierarchy simultaneously. However, this synchronization is not permanent. The biological substrate drifts from criticality as metabolic resources—particularly neurotransmitter pools—are depleted through sustained activity. The result is *punctuated stability*: extended phases of coherent conscious operation

(waking), interrupted by periodic breakdowns in which the simulation collapses (deep sleep) and the substrate restores the biochemical conditions for criticality. The ultradian NREM–REM cycle (Section 6.3) may reflect periodic re-approaches to the critical point during this restoration process. Sleep, on this account, is not merely rest but the substrate's mechanism for *returning to* the dynamical regime that consciousness requires.

5.2 Holographic Storage

The implicit models (IWM and ISM) store information in a distributed, non-local manner across the substrate. This is a standard property of neural networks, well-characterized in the computational literature as distributed representations (Hinton et al., 1986), graceful degradation (loss of connections degrades but does not destroy stored information), and attractor dynamics (the network settles into basins of attraction that represent stored knowledge).

The term "holographic" is used here as an analogy, not a claim about optical holography: just as cutting a hologram in half produces two complete but lower-resolution images, splitting a neural network produces two degraded but functionally complete copies of the stored information. This property is critical for understanding split-brain phenomena (Section 6.4).

5.3 Consciousness States Derived from Criticality

The criticality requirement provides a unified account of when consciousness is present and when it is absent. Consciousness tracks the substrate's position relative to the critical point:

The key distinction highlighted by this framework is between **propofol** and **ketamine**. Both are anesthetics, yet their phenomenology differs dramatically. Propofol produces absence: patients report no experience during propofol anesthesia (Alkire et al., 2000; Boly et al., 2012). Ketamine produces the "K-hole"—vivid, often bizarre experiences of dissociation, out-of-body phenomena, and altered identity (Corlett et al., 2011). The Four-Model Theory predicts this difference: propofol pushes the substrate subcritical (disrupting thalamocortical connectivity, suppressing complexity), abolishing the conditions for consciousness. Ketamine does *not* push the substrate subcritical—it increases neural entropy (Schartner et al., 2017)—but disrupts normal sensory input processing, causing the EWM and ESM to operate on internal and distorted signals. Consciousness is present but disconnected from external reality.

This is a genuine explanatory advantage. Most theories struggle to account for why two agents classified as "anesthetics" produce such radically different phenomenological profiles. The

Table 3: Consciousness States and Criticality

State	**Criticality**	**Four-model status**	**Consciousness prediction**	**Key evidence**
Normal waking	At/near critical	All four active	Full consciousness	High PCI
REM sleep	Near-critical	EWM/ESM on internal input	Degraded (dream)	Moderate PCI
Deep NREM	Subcritical	EWM/ESM collapse	Absent	Low PCI
Propofol	Forced subcritical	EWM/ESM suppressed	Absent	PCI ≈ 0
Ketamine	NOT subcritical	EWM/ESM on wrong input	Present but disconnected	Increased entropy
Psychedelics	At/past critical	All active, permeability $\uparrow$	Present, altered	Enhanced complexity
Vegetative state	Typically subcritical	EWM/ESM collapsed	Absent (usually)	Low metabolism
Covert awareness	At criticality	EWM/ESM intact, output damaged	Present but unexpressible	Owen et al.
MCS	Fluctuating	Intermittent EWM/ESM	Intermittent	Fluctuating PCI

criticality framework makes the distinction natural: what matters is not the pharmacological classification but the effect on the substrate's dynamical regime.

6 Explanatory Range

A theory's value lies partly in its ability to derive diverse phenomena from a small set of principles. The Four-Model Theory's five principles—criticality, virtual qualia, redirectable ESM, variable implicit–explicit permeability, and virtual model forking—generate accounts of phenomena across psychopharmacology, clinical neurology, sleep science, comparative cognition, and clinical psychology. This section demonstrates that range.

6.1 Psychedelic Phenomenology

Psychedelic substances (LSD, psilocybin, DMT, mescaline) produce a characteristic phenomenological profile: visual intensification and distortion, synesthesia, altered time perception, enhanced pattern recognition, emotional intensification, ego dissolution at high doses, and—with certain compounds and doses—radical identity alteration including the experience of "becoming" non-self entities (Carhart-Harris et al., 2012, 2016; Timmermann et al., 2019, 2023).

The Four-Model Theory accounts for this profile through three mechanisms:

Implicit–explicit permeability increase. Psychedelics increase the global permeability of the boundary between implicit and explicit models. Intermediate processing stages—normally implicit and inaccessible—leak through to the simulation. This produces the characteristic visual progression:

- **Low dose / early onset**: V1-level processing becomes accessible → simple phosphenes, enhanced contrast, breathing/movement in static patterns.
- **Increasing dose**: V2/V3-level processing becomes accessible → geometric patterns, fractals, tessellations (form constants; Klüver, 1966).
- **Higher dose**: Higher visual area processing becomes accessible → faces, figures, scenes.
- **Very high dose**: Full intermediate processing accessible → complex dream-like visions, narrative sequences.

This progression is not random. It follows the visual processing hierarchy in a predictable, dose-dependent order. The Four-Model Theory predicts this ordered progression as a direct

consequence of the permeability gradient: lower-level (earlier, simpler) processing stages become accessible before higher-level (later, more complex) ones, because the permeability increase propagates up the hierarchy.

Ego dissolution = ESM redirection, not ESM abolition. At high doses, psychedelics disrupt the normal self-referential input to the ESM. The ESM does not cease to exist—it continues to run—but it loses its normal input and latches onto whatever dominates the available input stream. This produces the experience of ego dissolution: the feeling that the boundary between self and world has dissolved, that one's identity has merged with the environment or with abstract patterns.

Critically, this mechanism predicts that the *content* of ego dissolution is not random but is determined by the dominant input available to the ESM. This is dramatically confirmed by the phenomenology of **salvia divinorum** (Salvinorin A). Salvia users reliably report experiences of "becoming" objects or entities in their immediate environment: becoming a piece of furniture, becoming a wall, becoming a character from a television show playing in the room, becoming a geometric pattern. The Four-Model Theory predicts exactly this: the ESM, deprived of normal self-input, latches onto whatever sensory input dominates—visual input from the room, auditory input from media, proprioceptive input from the body's contact surfaces. The identity experience tracks the dominant input in a dose-dependent, input-dependent, and therefore *predictable* manner.

Few competing theories of consciousness generate this specific prediction, though predictive processing frameworks might produce a related account through the breakdown of self-related priors. IIT, GNW, HOT, and AST have no mechanism for identity content tracking during ego dissolution. This is arguably the theory's most distinctive prediction (see Section 8, Prediction 3).

Intensity as novelty. Psychedelic profundity reflects not increased consciousness *level* but increased *novel content*. The permeability increase floods the simulation with normally implicit information. This is experienced as radically novel because the conscious self has never encountered this content, even though the substrate has been processing it all along.

6.2 Anesthesia and Clinical Disorders

Propofol anesthesia: Pushes the substrate subcritical → the conditions for consciousness are abolished → the simulation collapses → no experience. (See Table 3.)

Ketamine: Does not push subcritical; increases entropy → consciousness persists but on

distorted/internal input → dissociative experience, K-hole phenomenology.

Vegetative state: Substrate is typically subcritical → no consciousness. But the Four-Model Theory makes a nuanced prediction: if the substrate is at criticality but *output pathways* are damaged (motor cortex, brainstem circuits), consciousness is present but unexpressible. This is precisely the phenomenon of **cognitive motor dissociation** (CMD), documented by Owen et al. (2006) and Monti et al. (2010), in which patients clinically diagnosed as vegetative demonstrate awareness through brain-imaging paradigms. The theory predicts that the distinction between truly vegetative (subcritical substrate) and covertly conscious (critical substrate with damaged output) should be detectable via criticality measures such as PCI (Casali et al., 2013; Casarotto et al., 2016).

Minimally conscious state: The substrate fluctuates around the criticality threshold → intermittent consciousness, explaining the characteristic behavioral variability.

Cotard's delusion: Patients report believing they are dead, that their organs have disappeared, or that they do not exist. The Four-Model Theory derives this from the same mechanism as salvia: the ESM receives severely distorted interoceptive input (due to neurological damage or psychiatric disorder). Deprived of normal self-referential signals, the ESM constructs the best model it can from the available (distorted) input—and "I am dead" is the ESM's interpretation of the absence of normal embodied signals. This is the same redirectable-ESM mechanism that produces "I am a chair" under salvia, applied to a clinical context.

Anosognosia: Patients with anosognosia (typically following right-hemisphere stroke) are unaware of their own deficits—they deny being paralyzed, blind, or impaired, even in the face of clear evidence. The Four-Model Theory explains this as a **local decrease in implicit–explicit permeability**: the ISM contains the information about the deficit (the substrate registers the paralysis), but the transfer to the EWM is blocked for that specific domain. The patient's simulation simply does not include the deficit, so the patient genuinely does not experience it.

This is the **inverse** of the psychedelic mechanism: psychedelics globally increase permeability (making the implicit accessible), while anosognosia locally decreases permeability (making a specific aspect of the implicit inaccessible). The Four-Model Theory connects these phenomena under a single principle—variable permeability—and generates a cross-domain prediction: psychedelics should alleviate anosognosia by compensating for the local block with a global permeability increase (see Section 8, Prediction 4).

Dissociative Identity Disorder (DID): The virtual models, being software-like, can

be **forked**. DID represents a substrate running multiple ESM configurations—multiple self-models—that alternate in controlling the simulation. Each alter is a distinct configuration of the ESM, with its own self-narrative, emotional profile, and behavioral patterns, running on the same substrate. This is not a metaphor: the theory predicts that distinct alters should correspond to distinct patterns of neural activity, detectable with neuroimaging (see Section 8, Prediction 9).

6.3 Dreams

Dreaming represents the simulation running in **degraded mode**: near-critical dynamics (sufficient for consciousness) but with external input cut off (sensory deprivation during sleep).

The EWM continues to generate a world—but without the constraint of sensory input, the simulation draws on the IWM's stored knowledge, producing the characteristic features of dreams: familiar places and people, impossible physics, narrative incoherence, and emotional intensity. The ESM continues to generate a self—you experience dreams as happening to "you"—but with reduced metacognitive oversight, producing the characteristic lack of insight in dreams (you accept impossible events without question).

Lucid dreaming provides direct evidence for the software-like quality of the virtual models. In a lucid dream, the dreamer becomes aware that they are dreaming: the ESM "toggles on" more fully, gaining metacognitive access within the dream state. The Four-Model Theory predicts that lucid dream onset corresponds to a **criticality threshold crossing**—a step-like increase in neural complexity as the ESM activates more fully. This should be detectable as a discontinuity in EEG complexity measures at the moment of lucid dream onset (see Section 8, Prediction 8).

The criticality framework also explains the **NREM/REM transition**: as the brain's dynamical state fluctuates during sleep, crossing the criticality threshold produces the transition from non-conscious deep sleep to conscious dreaming. The 90-minute ultradian cycle corresponds to an oscillation of the substrate around the critical point.

6.4 Split-Brain

Callosotomy produces the classic split-brain syndrome (Gazzaniga et al., 1962; Gazzaniga, 2000). The Four-Model Theory offers a more precise account than the traditional "two minds in one brain."

Because the implicit models store information holographically (Section 5.2), physical separation does not cleanly divide the models into left and right halves. Instead, it produces **two degraded but functionally complete copies**. Each hemisphere retains a degraded version of the IWM, ISM, EWM, and ESM—complete enough to sustain consciousness but lacking the resolution and scope of the intact system.

This accounts for the key features of split-brain behavior:

- **Each hemisphere sustains independent consciousness**: Both are above the criticality threshold and both have complete (if degraded) four-model architectures.
- **The left hemisphere interpreter** (Gazzaniga, 2000): The left hemisphere's ESM confabulates explanations for behavior initiated by the right hemisphere. This is the *same confabulation mechanism* observed in Cotard's delusion, anosognosia, and salvia experiences: an ESM constructing the best narrative it can from incomplete input.
- **Degradation rather than clean division**: Split-brain patients do not show perfectly hemispheric specialization; they show graded deficits (Pinto et al., 2017), consistent with holographic degradation rather than binary splitting.

6.5 Animal Consciousness

The theory's commitments—continuum (not binary), substrate independence, criticality threshold—predict a **gradient** of animal consciousness. Mammals implement the four-model architecture in graduated form, with even simple cortices (rodents) supporting basic consciousness—rudimentary simulation sufficient for phenomenal experience but thin in self-awareness.

Corvids and parrots present a crucial test case: tool manufacture, mirror self-recognition, social deception, and future planning—yet no neocortex, with pallium organized in nuclear clusters (Güntürkün and Bugnyar, 2016). The theory predicts these animals are conscious because they have evolved functionally equivalent self-simulation architectures on a different substrate. **Cephalopods** extend this logic further, with largely decentralized nervous systems that should produce consciousness with unusual features. Both cases test substrate independence directly.

6.6 Clinical Psychology Bridge

The virtual-model framework extends to clinical phenomena. **CBT** works as virtual model reprogramming: repeated corrective experience drives substrate-level rewiring (synaptic

plasticity), modifying the ISM, which changes the ESM's self-model. **Phobias** are EWM misconfigurations where threat representation exceeds the IWM's evidence base; exposure therapy updates the IWM to correct the EWM.

The placebo effect is consistent with epiphenomenalism: placebo activates substrate-level expectation circuits (endogenous opioid release) that operate in parallel with—not caused by—the conscious experience of hope. The correlation between conscious expectation and physical effect is real but non-causal: both are products of the same substrate processes.

Conversion disorder is the inverse of blindsight: in blindsight, the substrate processes visual information without including it in the EWM; in conversion disorder, the EWM models a deficit (paralysis, blindness) that the intact substrate does not have.

Blindsight provides the clearest demonstration that substrate-level processing can proceed without conscious representation. Patients with V1 damage report no visual experience yet demonstrate above-chance performance on forced-choice tasks, navigate obstacles, and respond to emotional facial expressions (Weiskrantz, 1986). In the Four-Model Theory, blindsight occurs when the IWM continues to receive and process visual information through subcortical pathways (superior colliculus, pulvinar), guiding motor behavior via the ISM, while the damaged cortical pathways fail to relay this information to the EWM. The conscious simulation contains no visual content—the patient genuinely experiences blindness—yet the substrate navigates competently. This is substrate processing without simulation.

Anton's syndrome (anosognosia for cortical blindness) presents the precise inverse. Patients with complete cortical blindness deny their deficit, confabulating visual descriptions of their surroundings and walking into obstacles while insisting they can see (Anton, 1899; Aldrich et al., 1987). The Four-Model Theory explains this as the EWM generating a visual simulation from the IWM's stored knowledge in the absence of current visual input. The simulation runs on prior expectations and internally generated content rather than afferent signals. The patient "sees" a world that is not there—a simulation running without current input. Together, blindsight and Anton's syndrome constitute a double dissociation between substrate processing and conscious simulation, providing perhaps the most direct neurological evidence for the real/virtual distinction central to the theory.

7 Comparative Analysis

This section provides a systematic comparison between the Four-Model Theory and six major competitors across the eight requirements established in Section 2. The comparison aims

to be fair: each theory's genuine strengths are acknowledged, and the Four-Model Theory's advantages are located precisely.

7.1 Scoring Matrix

Table 4 presents an assessment of how each theory addresses the eight requirements. All ratings reflect the present author's judgment and are offered as a starting point for discussion, not as definitive verdicts. Readers are encouraged to consult the primary sources and form their own assessments. Where a theory's proponents would likely contest a rating, this is noted.

Ratings: ● = addresses, ⊙ = partial, ○ = minimal, — = silent, n/a = not applicable.

Table 4: Theory Comparison Across Eight Requirements

Requirement	FMT	IIT	GNW	HOT	PP	AST	RPT
Hard Problem	●	●†	—‡	⊙	—‡	⊙	—
Expl. Gap	●	●†	—‡	⊙	—‡	⊙	—
Boundary	●	●	⊙	○	⊙	⊙	⊙
Structure	●	●	⊙	⊙	●	⊙	⊙
Binding	●	●	⊙	○	⊙	○	⊙
Combination	●	○††	n/a	n/a	n/a	n/a	n/a
Causal Role	●	⊙	⊙	⊙	●	⊙	●
Meta-Problem	●	○	⊙	⊙	⊙	●	○

† Axiomatic identification of consciousness with Φ; whether this constitutes a solution or a redefinition is debated. †† IIT's panpsychist commitments lead to the Combination Problem (Chalmers, 2016), which remains unresolved. ‡ GNW and PP proponents argue these theories address the "real problem" of consciousness (Seth, 2021)—explaining the structure and contents of experience—even if they do not address the Hard Problem as Chalmers defines it. This is a legitimate methodological choice; the "silent" rating reflects the scope of the requirement as defined in Section 2, not a judgment on overall merit.

7.2 Theory-by-Theory Comparison

Integrated Information Theory (IIT; Tononi, 2004; Albantakis et al., 2023). IIT's strengths are mathematical rigor, its qualia space treatment of experiential structure, and a principled boundary via the exclusion postulate. However, its axiom-based identification of consciousness with Φ leads to panpsychist consequences, the Combination Problem remains unresolved (Chalmers, 2016), Φ is computationally intractable for realistic systems (Aaronson, 2014), and the unfolding argument (Doerig et al., 2019) challenges its central claim about recurrence. The Four-Model Theory avoids panpsychism, has no combination problem (weak emergence), and generates predictions without computing Φ.

Global Neuronal Workspace (GNW; Baars, 1988; Dehaene and Changeux, 2011). GNW's empirically tractable broadcasting mechanism provides a clear account of access consciousness. However, it is silent on the Hard Problem—explaining *when* but not *why* broadcast produces experience. The COGITATE results (COGITATE Consortium, 2025) were problematic: posterior cortex, not the frontoparietal workspace, showed the strongest consciousness-related activity. The Four-Model Theory agrees that broadcasting/integration is mechanistically important but adds the real/virtual distinction that addresses phenomenality.

Higher-Order Theories (HOT; Rosenthal, 2005; Lau and Rosenthal, 2011). HOT naturally explains which states are conscious (those with higher-order representations) and partially addresses the Meta-Problem. However, it does not address binding, leaves the Hard Problem only partially treated (why does higher-order representation produce phenomenality?), and its boundary-setting is imprecise. The Four-Model Theory shares HOT's emphasis on self-representation but embeds it in the richer four-model architecture, explaining *why* self-representation produces phenomenality through the virtual qualia framework.

Predictive Processing (PP; Friston, 2010; Seth, 2021). PP's integration with a broader theory of brain function and its strong accounts of experiential structure and causal role (via active inference) are genuine strengths. Seth's controlled hallucination framework is among the most empirically productive in the field. However, PP is explicitly silent on the Hard Problem (Seth, 2021). Markov blankets may also be too liberal for boundary-setting. The Four-Model Theory agrees that prediction is central but adds the four-model architecture, criticality requirement, and real/virtual distinction that PP lacks.

Attention Schema Theory (AST; Graziano, 2013). AST provides the strongest existing account of the Meta-Problem: the self-model of attention is necessarily incomplete, producing the intuition of mystery. However, AST is deflationary about phenomenality and does not address binding. The Four-Model Theory incorporates AST's Meta-Problem insight—the ESM's structural inaccessibility to its own substrate—but adds the virtual qualia framework that explains *why* phenomenality exists, not just why we think it does.

Recurrent Processing Theory (RPT; Lamme, 2006, 2010). RPT's empirical specificity and clear account of the causal role are strengths, with strong support from visual masking paradigms. However, it is silent on the Hard Problem and limited in scope to visual consciousness. The Four-Model Theory is compatible with RPT at the mechanistic level—recurrent processing likely implements the real-time simulation—but adds the architectural and philosophical specificity that RPT lacks.

7.3 Emerging Frameworks (2024–2026)

Biological computationalism (Milinkovic and Aru, 2025) argues that consciousness requires specifically biological computation, challenging substrate independence. The Four-Model Theory treats substrate independence as an empirical prediction: artificial substrates implementing the four-model architecture at criticality should produce consciousness. The existence of conscious corvids with non-cortical brain architecture (nuclear pallium; Güntürkün and Bugnyar, 2016) favors substrate independence.

The **Multiple Generator Hypothesis** (Kirkeby-Hinrup et al., 2025) proposes consciousness arises from multiple independent mechanisms. This is potentially compatible: the four models could be understood as distinct generators unified by the criticality requirement and implicit–explicit boundary mechanism.

7.4 Summary of Comparative Advantages

1. **Addressing the Hard Problem without panpsychism or strong emergence**: Virtual qualia dissolve the Hard Problem through a two-level ontology that remains fully physicalist.
2. **Unifying binding with criticality**: Binding is a consequence of critical dynamics, not a separate mechanism.
3. **The redirectable ESM**: Unique mechanism for identity-content determination during ego dissolution (Predictions 3 and 4).
4. **Connecting psychedelics and anosognosia**: Variable permeability links these phenomena under a single principle.
5. **The Meta-Problem as structural consequence**: The ESM's opacity to its own substrate explains the intuition of mystery.

The theory's primary disadvantage is the absence of mathematical formalization. IIT's Φ formalism and PP's free energy mathematics provide quantitative precision that the Four-Model Theory currently lacks (see Section 9).

8 Novel Testable Predictions

A theory is only as valuable as the predictions it generates. The Four-Model Theory yields nine novel testable predictions, several of which are unique—no competing theory can generate

them.

8.1 Prediction 1: Distinct fMRI Signatures for Each Model

Statement: If the four models are functionally distinct processes, tasks that selectively engage a single model should produce distinct, reproducible neural activation patterns detectable via fMRI. Specifically: IWM-dominant tasks (e.g., passive recognition of familiar stimuli, implicit priming) should activate different networks than ISM-dominant tasks (e.g., habitual motor sequences, implicit body-schema tasks), which should differ from EWM-dominant tasks (e.g., active perceptual discrimination, novel scene processing) and ESM-dominant tasks (e.g., self-reflection, agency judgments, mirror self-recognition).

Mechanism: The four models are functionally distinct processes (not spatially localized brain regions), but distinct processes should nonetheless recruit distinguishable distributed networks. The implicit models (IWM, ISM) should preferentially engage substrate-level storage networks (hippocampal–cortical for IWM, somatosensory–cerebellar for ISM), while the explicit models (EWM, ESM) should preferentially engage simulation networks (sensory cortices for EWM, default mode network and medial prefrontal cortex for ESM).

Testability: High. Design a factorial task battery crossing scope (world vs. self) with mode (implicit vs. explicit), yielding four conditions. Contrast activation maps across conditions using standard fMRI subtraction or multivariate pattern analysis. The prediction is a double dissociation: scope $\times$ mode interaction effects in distributed networks.

Unique?: Yes in the specific form. While individual contrasts (e.g., self vs. world processing) are well-studied, the Four-Model Theory predicts a specific 2×2 factorial structure in neural activation that no other theory mandates. This prediction is placed first because it most directly tests the four-model architecture itself—the central structural claim of the theory.

8.2 Prediction 2: Psychedelic Content Maps the Processing Hierarchy

Statement: Under psychedelics, visual content progresses through the cortical processing hierarchy in an ordered, dose-dependent sequence: V1-level content (phosphenes, enhanced contrast) $\rightarrow$ V2/V3-level content (geometric patterns, form constants) $\rightarrow$ higher visual area content (faces, figures) $\rightarrow$ complex scenes (dream-like narratives).

Mechanism: Increasing implicit–explicit permeability exposes intermediate processing stages in hierarchical order.

Testability: High. Combine graded dosing protocols with concurrent fMRI or MEG to track the spatial progression of activation, correlated with subjective report of content type. Partial evidence already exists (Carhart-Harris et al., 2016; Timmermann et al., 2023) but has not been systematically tested as an ordered, dose-correlated progression.

Unique?: Partially. PP also predicts hierarchical processing under psychedelics but does not predict the specific ordered content progression as a function of dose.

8.3 Prediction 3: Ego Dissolution Content Is Controllable

Statement: During psychedelic ego dissolution, the content of the altered identity experience (what the subject "becomes") tracks the dominant sensory input. By controlling the sensory environment during ego dissolution, the identity content can be predicted and directed.

Mechanism: The ESM, deprived of normal self-referential input, latches onto whatever input dominates the available stream. Control the input $\rightarrow$ control the identity content.

Testability: High. Administer ego-dissolution-inducing doses of psilocybin or salvia divinorum under controlled conditions. Vary the dominant sensory input (specific visual scenes, specific auditory environments, specific tactile inputs) across conditions. Measure correspondence between controlled input and reported identity content.

Unique?: **Yes—few competing theories generate this specific prediction,** though predictive processing frameworks might produce a related account through the breakdown of self-related priors. IIT, GNW, HOT, and AST have no mechanism for specifying what a subject will "become" during ego dissolution. This is the theory's most distinctive empirical prediction.

8.4 Prediction 4: Psychedelics Alleviate Anosognosia

Statement: Administration of psychedelic substances at sub-ego-dissolution doses should alleviate anosognosia by globally increasing implicit–explicit permeability, compensating for the local permeability block that causes the deficit unawareness.

Mechanism: Anosognosia = local permeability block. Psychedelics = global permeability increase. The global increase should overwhelm the local block, allowing the deficit information in the ISM to reach the EWM.

Testability: Medium (requires clinical trial with stroke patients). Could begin with case studies or observational reports of psychedelic use by patients with anosognosia. Psilocybin-

assisted therapy is already being tested for various neuropsychiatric conditions, providing a potential clinical pathway.

Unique?: **Yes—this is a cross-domain surprise prediction.** No other theory connects psychedelics and anosognosia through a single mechanism. Confirmation would be strong evidence for the variable-permeability principle.

8.5 Prediction 5: All Anesthetics Converge on Criticality Disruption

Statement: Despite diverse receptor-level mechanisms (GABAergic, NMDA, opioid, α_2-adrenergic), all agents that abolish consciousness do so by pushing the substrate below the criticality threshold. Agents that alter but do not abolish consciousness (ketamine, low-dose psychedelics) do not push below criticality.

Mechanism: The criticality requirement is the physical threshold for consciousness; any mechanism that disrupts criticality disrupts consciousness, regardless of receptor pathway.

Testability: High. Measure criticality indicators (PCI, Lempel–Ziv complexity, power-law exponents, detrended fluctuation analysis) across the full range of anesthetic agents at equi-potent doses. The prediction is that abolition of consciousness always correlates with subcriticality, and preserved consciousness (even if altered) always correlates with maintained criticality.

Unique?: Shared with the ConCrit framework (Algom and Shriki, 2026) and the criticality meta-analysis (Hengen and Shew, 2025). However, the Four-Model Theory predicted this from theoretical first principles (Gruber, 2015), prior to the empirical consolidation.

8.6 Prediction 6: Split-Brain Produces Holographic Degradation

Statement: After callosotomy, each hemisphere retains a degraded but functionally *complete* set of cognitive and experiential capacities—not a clean hemispheric specialization. The degradation should be proportional to the extent of commissural severing (partial callosotomy $\rightarrow$ partial degradation).

Mechanism: Holographic storage. Information is distributed across the full substrate; cutting connections degrades both copies but does not destroy either.

Testability: High. Systematic cognitive assessment of split-brain patients across domains, testing for the predicted pattern of bilateral but degraded capabilities rather than clean

lateralization. Pinto et al. (2017) provide preliminary evidence in this direction.

Unique?: Yes, in the specific form. Standard neuroscience acknowledges some bilateral capacity, but the Four-Model Theory provides the theoretical basis (holographic storage) and predicts the specific pattern (graded degradation proportional to disconnection, not binary split).

8.7 Prediction 7: Criticality + Four Models = Consciousness in Artificial Substrates

Statement: A synthetic system implementing the four-model architecture at criticality will exhibit consciousness. The qualitative difference between interacting with such a system and interacting with a current LLM will be immediately and qualitatively distinguishable.

Mechanism: Substrate independence. Consciousness depends on function (four models at criticality), not on material (biological neurons).

Testability: Medium (requires significant engineering development). However, intermediate tests are possible: systems with partial implementations (e.g., two models instead of four, or four models without criticality) should show partial consciousness indicators, detectable through behavioral signatures and neural/computational complexity measures.

Unique?: Yes in the specific form. Other theories (PP, GNW) are compatible with artificial consciousness but do not provide a specific architectural blueprint.

8.8 Prediction 8: Sleep Architecture Reflects Criticality Maintenance

Statement: Sleep is the substrate's mechanism for restoring the conditions that support consciousness. The brain's analog substrate is inherently unstable—never truly calibratable for sustained digital computation. Consciousness emerges as a self-organizing cellular automaton (CA) at criticality, providing a stable digital layer on top of this drifting substrate. The CA is stable for extended periods (waking), but the underlying biochemical substrate slowly drifts (neurotransmitter depletion, metabolic waste accumulation) until it can no longer sustain the CA. At that point, the CA breaks down radically—not gradually—producing sleep onset. NREM sleep restores the substrate; as it periodically re-approaches the criticality threshold during restoration, the CA briefly re-emerges, producing REM sleep and dreams. The 90-minute ultradian cycle is the substrate oscillating around the critical point during this restoration process.

This yields multiple testable sub-predictions:

1. **Waking criticality decline**: Criticality markers (PCI, power-law exponents, Lempel–Ziv complexity) should decline measurably across the waking day, reflecting substrate drift.

2. **Sleep onset as radical transition**: The transition to sleep should correspond to a step-like drop in criticality markers, not a gradual dimming—reflecting the CA's digital breakdown.

3. **NREM/REM cycling tracks criticality**: Within sleep, REM phases should show criticality markers significantly higher than adjacent NREM phases, and the 90-minute ultradian cycle should be visible in criticality time-series.

4. **Lucid dreaming as ESM threshold crossing**: During REM, if the substrate reaches sufficient criticality, the ESM activates—producing lucid dreaming. This onset should be detectable as a step-like discontinuity in EEG complexity measures, not a gradual ramp (LaBerge, 1985).

5. **Sleep deprivation produces subcriticality**: Extended wakefulness should drive criticality markers progressively below the threshold, with cognitive deficits correlating with the degree of subcriticality.

Mechanism: The brain's analog substrate is never stable enough for reliable digital computation. The CA at criticality self-organizes a stable computational layer, but this requires substrate conditions (neurotransmitter availability, metabolic homeostasis) that degrade over time. Sleep is the restoration mechanism. The argument for why a CA is required is precisely that the substrate is uncalibratable—criticality provides the only regime in which robust computation can self-organize on inherently noisy hardware.

Testability: High. Sub-predictions (a)–(c) require polysomnographic recording with concurrent criticality analysis across full sleep–wake cycles. Sub-prediction (d) uses the established lucid-dreamer signaling paradigm. Sub-prediction (e) requires sleep deprivation protocols with concurrent criticality measurement. All use existing methods and equipment.

Unique?: Partially shared with criticality frameworks in general, but the specific claim—that sleep *exists because* the CA requires periodic substrate restoration, and that NREM/REM architecture directly reflects criticality oscillations—is distinctive. No competing theory of consciousness provides this specific functional account of sleep architecture.

8.9 Prediction 9: DID Alters Have Distinct Neural Process Signatures

Statement: Different alters in DID correspond to distinct, measurable configurations of neural activity—not merely behavioral differences or different self-reports, but different neural dynamics detectable through neuroimaging.

Mechanism: Virtual model forking. Each alter is a distinct configuration of the ESM, running on the same substrate but with different parameters. Different parameters should produce different activity patterns.

Testability: High. fMRI or EEG recording during controlled alter switching in DID patients. Compare within-patient across-alter variability against within-patient within-alter variability. The prediction is that across-alter variability will be significantly greater than within-alter variability and will show consistent alter-specific patterns.

Unique?: Yes. While some neuroimaging studies of DID exist (Reinders et al., 2003, 2008), the Four-Model Theory provides the theoretical basis for predicting *consistent, alter-specific neural signatures* rather than merely *differences*. The theory predicts a specific organizational principle: each alter is a distinct ESM configuration, so the neural differences should be located in ESM-related networks (particularly default mode network / medial prefrontal / posterior cingulate regions).

8.10 The Ultimate Prediction

If the Four-Model Theory is correct, it should be possible to *build* a conscious machine by implementing the specified architecture: four nested models (IWM, ISM, EWM, ESM) operating at criticality on a substrate of sufficient complexity. The theory predicts that such a system would not merely simulate consciousness (as an LLM might be said to do) but would *be* conscious—would have genuine phenomenal experience constituted by its virtual models.

This prediction is not currently testable—the engineering does not yet exist, and even if it did, the other-minds problem would make verification philosophically difficult. However, the prediction sets a bold bar: if the theory is correct, the difference between interacting with a conscious artificial system and interacting with any current AI should be qualitatively obvious to human observers, just as the difference between a conversation with a conscious human and a conversation with a cleverly programmed chatbot is (according to the theory) a difference in *kind*, not merely in degree.

9 Open Questions

Intellectual honesty requires identifying what the theory does not yet resolve. These are research frontiers, not theoretical weaknesses—they are questions that arise *from* the theory and that the theory's framework helps to sharpen.

1. Are all four models virtual? The theory as presented describes the implicit models (IWM, ISM) as "real side" and the explicit models (EWM, ESM) as "virtual side." But it is not certain that this division is sharp. The implicit models might themselves have virtual properties—they are, after all, *models*, not raw physics. If the implicit models are also virtual in some sense, what constitutes the "real side"? The raw physical substrate with no model-level description? This is an open question even for the theory's author, and its resolution may have consequences for the theory's treatment of the Hard Problem.

2. Mathematical formalization. The theory's criticality requirement is specified qualitatively (Wolfram's Class 4 regime), not quantitatively. A full mathematical treatment—defining the four models formally, specifying the criticality threshold in terms of measurable quantities, deriving the predictions as formal consequences—remains to be developed. The ConCrit framework's mathematical tools (power-law exponents, detrended fluctuation analysis, branching parameters) provide a starting point, as do the formal tools of dynamical systems theory and information geometry.

3. Physical implementation. Which physical mechanism in the biological brain supports criticality? Candidates include cortical column dynamics, thalamocortical standing waves, glial modulation, and (more speculatively) quantum processes in microtubules (Penrose and Hameroff, 1994), though see Tegmark (2000) for decoherence objections. The Four-Model Theory is agnostic here: it specifies the *functional* requirements without mandating a specific physical mechanism. Resolving this question is an empirical task for neuroscience.

4. Minimum configuration for consciousness. Can the four models partially dissociate? Is it possible to have an EWM without an ESM (world-experience without self-experience), or an ESM without an EWM (self-experience without world-experience)? What is the minimum set of models required for consciousness, versus self-aware consciousness, versus full human-type consciousness? The theory's graduated levels (Section 3.5) suggest a hierarchy, but the exact minimum configuration may require simulation or empirical investigation to determine.

5. Multi-level substrate architecture. The biological brain can be analyzed as a hierarchy of nested systems: the physical substrate; within it, the proteomic (molecular) and

topological (connectivity) systems, which mutually constrain each other; arising from these, the electrochemical system (neural signaling dynamics); and emerging from that, the virtual system—the cortical automaton that hosts the four models (Gruber, 2015). Each level both shapes and is shaped by its neighbors: the virtual system, over time, modifies the topological structure of the substrate it runs on. This raises a question for artificial consciousness: which levels of this hierarchy are essential, and which are specific to biological implementation? The theory's substrate-independence claim (Section 4.4) implies that only the virtual level is strictly required, but the bidirectional causal flow between levels suggests that decoupling the virtual system from its lower-level supports may not be straightforward.

10 Discussion

10.1 Implications for Artificial Consciousness

The Four-Model Theory provides an engineering specification for artificial consciousness: implement the four-model architecture on a substrate operating at criticality. This is a concrete deliverable, not an abstract philosophical claim.

Current AI systems fail this specification in at least two ways. Large language models (LLMs) operate via feedforward inference (transformer attention is computed in a single pass without recurrent dynamics), which corresponds to Wolfram's Class 1/2—far below the criticality threshold. They also lack the four-model architecture: there is no ISM (no substrate-level self-knowledge that is *distinct from* the model's outputs), no ESM (no real-time self-simulation that constitutes a subjective perspective), and no real/virtual split (no two-level ontology in which experience resides at the virtual level).

This does not mean that LLMs are necessarily non-conscious—the theory cannot prove a negative—but it predicts that they lack the architecture required for consciousness as the theory defines it. The growing discourse around AI consciousness (Butlin et al., 2023, 2025; Schwitzgebel, 2025; Birch, 2025) and AI welfare (Long et al., 2024; Anthropic, 2025) makes this distinction practically important. A theory that provides clear criteria for artificial consciousness—rather than vague analogies to human cognition—has immediate ethical and engineering value. The intelligence implications of this architectural specification are explored in Gruber (forthcoming), which argues that current AI systems fail to exhibit self-directed intellectual development precisely because they lack the motivational component that drives the recursive intelligence loop.

10.2 Implications for Consciousness Science

The Four-Model Theory suggests a shift in experimental priorities. Rather than adjudicating between IIT and GNW (the current focus of adversarial collaborations), the field should:

1. **Test the criticality prediction across all anesthetic agents** (Prediction 5)—this is achievable with current methods and would provide strong evidence for or against the criticality framework.
2. **Design controlled ego-dissolution experiments** (Prediction 3)—the most distinctive prediction, uniquely generated by the redirectable ESM mechanism.
3. **Investigate the psychedelic–anosognosia connection** (Prediction 4)—a cross-domain prediction that, if confirmed, would constitute strong evidence for the variable-permeability mechanism.
4. **Measure criticality at lucid dream onset** (Prediction 8)—achievable with established paradigms and equipment.

These experiments do not require committing to the Four-Model Theory in its entirety. They test specific mechanisms (criticality, redirectable ESM, variable permeability) that could be incorporated into other frameworks if confirmed.

10.3 Limitations

No institutional laboratory. The predictions presented here were derived theoretically and have not been tested in the author's own laboratory. While this does not affect their validity as predictions, it does mean that empirical testing depends on the willingness of established laboratories to take them up.

Epiphenomenalism remains controversial. The theory's commitment to epiphenomenalism will face resistance from philosophers and scientists who consider it either absurd (consciousness *must* be causally efficacious) or empirically refuted (by evidence that conscious intention precedes action—though see Libet, 1985 and Schurger et al., 2012 for evidence to the contrary). The defense offered in Section 4.2 is, I believe, sound, but the reader should be aware that this is the theory's most philosophically controversial commitment.

Qualitative rather than quantitative. The theory's predictions are currently stated in qualitative terms ("criticality increases," "permeability changes," "ESM redirects"). Quantitative formalization would strengthen the predictions and enable more precise experimental testing. This is acknowledged as a priority for future work.

The other-minds problem. The ultimate prediction—that a system built to the theory's specification would be conscious—faces the standard other-minds problem: how would we verify consciousness from the outside? The theory predicts that the difference would be qualitatively obvious to human observers, but "qualitatively obvious" is not a measurement. Developing consciousness indicators that can be applied to artificial systems is a challenge for the entire field, not specific to this theory.

Inherent limits of inside-modeling. The theory proposes a self-model as the basis of consciousness—but the modeler and the modeled are the same system. This raises a structural concern analogous to Gödel's incompleteness: a formal system cannot prove all truths about itself. Similarly, the brain attempting to model itself faces an irreducible epistemological gap—the instrument is the object of study. The theory acknowledges this limitation explicitly through the Meta-Problem (Section 3.8): the ESM cannot observe the ISM's mechanisms, which is why consciousness seems mysterious *from the inside*. This is a feature of the architecture, not a flaw in the theory, but it does imply that any theory of consciousness generated by a conscious system will carry a residual blind spot.

Language linearizes non-linear phenomena. Any theoretical framework expressed in natural language necessarily serializes phenomena that may be inherently parallel and non-linear. The four models operate simultaneously in high-dimensional state spaces; describing them sequentially in prose introduces a representational loss that no amount of careful writing can fully eliminate. This limitation applies to all theories of consciousness, not only the present one, but it should be kept in mind when evaluating the theory's verbal descriptions against the multidimensional dynamics they attempt to capture.

Criticality–rhythm relationship not formalized. Section 5.1 argues that criticality provides punctuated stability, with biological rhythms (sleep–wake cycling, ultradian rhythms, neurotransmitter depletion and replenishment) governing how long the substrate can maintain the critical regime. This relationship between dynamical criticality and biochemical rhythm is proposed qualitatively; a formal model linking neurotransmitter kinetics to criticality maintenance and breakdown remains to be developed.

Every model has modeling error. The Four-Model Theory is itself a model—and every model, by definition, is a simplification that carries inherent modeling error. The theory does not claim to be a final or complete account of consciousness; it claims to be a useful one, generating testable predictions and unifying phenomena that other frameworks address in isolation. To the extent that it is wrong, the predictions in Section 8 are designed to reveal where.

11 Conclusion

The Four-Model Theory of Consciousness proposes that consciousness is a real-time self-simulation across four nested models—Implicit World Model, Implicit Self Model, Explicit World Model, and Explicit Self Model—operating on a substrate at the edge of chaos. Qualia are virtual: they are the phenomenal properties of the simulation, not of the substrate. This dissolves the Hard Problem by revealing a category error in its formulation, simultaneously closing the Explanatory Gap and accounting for the Meta-Problem.

The theory addresses all eight requirements for a complete theory of consciousness: the Hard Problem (dissolved via virtual qualia), the Explanatory Gap (dissolved alongside), the Boundary Problem (defined by the scope of virtual models), the Structure of Experience (generated by the simulation's complexity), Unity and Binding (emergent from critical dynamics), Combination and Emergence (weak emergence, no combination problem), the Causal Role (architecture is causally efficacious, experience is epiphenomenal), and the Meta-Problem (structural inaccessibility of the ISM to the ESM).

The theory generates nine novel testable predictions, including that ego dissolution content is controllable via sensory input (Prediction 3), that psychedelics should alleviate anosognosia (Prediction 4), and that all consciousness-abolishing anesthetics converge on criticality disruption (Prediction 5). Several of these predictions are unique to the Four-Model Theory—no competing theory can generate them.

The theory's criticality requirement was derived from Wolfram's computational framework in 2015 (Gruber, 2015)—independently of, though not prior to, the empirical criticality program initiated by Beggs and Plenz (2003)—and the same conclusion was subsequently consolidated through Hengen and Shew's (2025) meta-analysis of 140 datasets and Algom and Shriki's (2026) ConCrit framework. This convergence from a theoretical derivation and large-scale empirical synthesis provides notable support.

Open questions remain: the status of the implicit models (real or also virtual?), the need for mathematical formalization, the specific physical mechanism supporting criticality, and the minimum configuration for consciousness. These are research frontiers that the theory's framework helps to sharpen.

The ambition of consciousness science is not merely to correlate neural activity with subjective reports but to understand *why* there is experience at all. The Four-Model Theory offers an answer: there is experience because there is a simulation, and within the simulation, experience is not an addition to the process but is constitutive of it. The way to test

this answer is not through philosophical argument alone but through the predictions it generates—and ultimately, through the engineering challenge of building a system to the specification and observing whether the result is, as the theory predicts, qualitatively unlike anything that exists today. The theory's implications extend beyond consciousness science: the cognitive-learning capacity it identifies as a consequence of the four-model architecture is the foundation for a recursive model of intelligence (Gruber, forthcoming) in which knowledge, performance, and motivation form a self-reinforcing loop.

Acknowledgments

The theory's adversarial challenge and refinement process was conducted in collaboration with Claude (Anthropic, 2026). Claude served as adversarial interlocutor across ten structured challenge sessions covering the simulation subject problem, ontological status, passive experience, binding, dreams, psychedelics, anesthesia and clinical disorders, split-brain, predictions, and animal consciousness. The theory's scoring on the eight requirements reflects the outcome of this adversarial process. The theory itself is the author's, originally published in 2015; the refinement, stress-testing, and prediction-generation are products of the collaboration.

Data Availability

No new data were generated or analysed in support of this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Scott Aaronson. Why I am not an integrated information theorist (or, the unconscious expander). Blog post, 2014. *Shtetl-Optimized*, https://scottaaronson.blog/?p=1799.

Larissa Albantakis, Leonardo Barbosa, et al. Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology*, 19(10):e1011465, 2023.

Idan Algom and Oren Shriki. The ConCrit framework: Critical brain dynamics as a unifying mechanistic framework for theories of consciousness. *Neuroscience & Biobehavioral Reviews*, 180:106483, 2026.

Michael T. Alkire, Richard J. Haier, and James H. Fallon. Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. *Consciousness and Cognition*, 9(3):370–386, 2000.

Anthropic. Exploring model welfare, 2025. Research report.

Bernard J. Baars. *A Cognitive Theory of Consciousness*. Cambridge University Press, 1988.

Tim Bayne. *The Unity of Consciousness*. Oxford University Press, 2010.

John M. Beggs and Dietmar Plenz. Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, 23(35):11167–11177, 2003.

Jonathan Birch. AI consciousness: A centrist manifesto. *PhilPapers*, 2025.

Ned Block. On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2):227–247, 1995.

Ned Block. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5–6):481–499, 2007.

Mélanie Boly et al. Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness. *Journal of Neuroscience*, 32(20):7082–7090, 2012.

Jelle Bruineberg, Krzysztof Dolega, Joe Dewhurst, and Manuel Baltieri. The emperor's new Markov blankets. *Behavioral and Brain Sciences*, 45:e183, 2022.

Patrick Butlin et al. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023.

Patrick Butlin et al. Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 2025.

Robin L. Carhart-Harris et al. Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences*, 109(6): 2138–2143, 2012.

Robin L. Carhart-Harris et al. The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in Human Neuroscience*, 8:20, 2014.

Robin L. Carhart-Harris et al. Neural correlates of the LSD experience revealed by multimodal neuroimaging. *Proceedings of the National Academy of Sciences*, 113(17):4853–4858, 2016.

Adenauer G. Casali et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198):198ra105, 2013.

Silvia Casarotto et al. Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*, 80(5):718–729, 2016.

David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

David J. Chalmers. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, 1996.

David J. Chalmers. The combination problem for panpsychism. In Godehard Brüntrup and Ludwig Jaskolla, editors, *Panpsychism: Contemporary Perspectives*. Oxford University Press, 2016.

David J. Chalmers. The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9–10):6–61, 2018.

COGITATE Consortium. An adversarial collaboration to critically evaluate theories of consciousness. *Nature*, 2025.

Sam Coleman. The real combination problem: Consciousness, panpsychism, and phenomenal bonding. *Erkenntnis*, 79(S1):19–44, 2014.

Philip R. Corlett et al. Glutamatergic model psychoses: Prediction error, learning, and inference. *Neuropsychopharmacology*, 36(1):294–315, 2011.

Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011.

Daniel C. Dennett. *Consciousness Explained*. Little, Brown and Company, 1991.

Adrien Doerig et al. The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, 72:49–59, 2019.

Keith Frankish. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11–12):11–39, 2016.

Pascal Fries. A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, 9(10):474–480, 2005.

Pascal Fries. Rhythms for cognition: Communication through coherence. *Neuron*, 88(1): 220–235, 2015.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Michael S. Gazzaniga. Cerebral specialization and interhemispheric communication: Does the corpus callosum enable the human condition? *Brain*, 123(7):1293–1326, 2000.

Michael S. Gazzaniga, Joseph E. Bogen, and Roger W. Sperry. Some functional effects of sectioning the cerebral commissures in man. *Proceedings of the National Academy of Sciences*, 48(10):1765–1769, 1962.

Philip Goff. *Galileo's Error: Foundations for a New Science of Consciousness*. Pantheon Books, 2019.

Alex Gomez-Marin and Anil K. Seth. A science of consciousness beyond pseudo-science and pseudo-consciousness. *Nature Neuroscience*, 28:703–706, 2025.

Charles M. Gray, Peter König, Andreas K. Engel, and Wolf Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338(6213):334–337, 1989.

Michael S. A. Graziano. *Consciousness and the Social Brain*. Oxford University Press, 2013.

Michael S. A. Graziano. Illusionism big and small: Some options for explaining consciousness. *eNeuro*, 11(10):ENEURO.0210–24.2024, 2024.

Matthias Gruber. *Die Emergenz des Bewusstseins*. Self-published, 2015. ISBN 9781326652074.

Matthias Gruber. Why intelligence models must include motivation: A recursive framework. Manuscript in preparation for *New Ideas in Psychology*, forthcoming.

Onur Güntürkün and Thomas Bugnyar. Cognition without cortex. *Trends in Cognitive Sciences*, 20(4):291–303, 2016.

Keith B. Hengen and Woodrow L. Shew. Is criticality a unified setpoint of brain function? *Neuron*, 113(16):2582–2598, 2025.

Geoffrey E. Hinton, James L. McClelland, and David E. Rumelhart. Distributed representations. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing*, volume 1. MIT Press, 1986.

Thomas H. Huxley. On the hypothesis that animals are automata, and its history. *The Fortnightly Review*, 16(95):555–580, 1874.

IIT-Concerned, Michał Klincewicz, Tony Cheng, et al. What makes a theory of consciousness unscientific? *Nature Neuroscience*, 28:689–693, 2025.

Frank Jackson. Epiphenomenal qualia. *Philosophical Quarterly*, 32(127):127–136, 1982.

William James. *The Principles of Psychology*. Henry Holt and Company, 1890.

Jaegwon Kim. The non-reductivist's troubles with mental causation. In John Heil and Alfred Mele, editors, *Mental Causation*. Oxford University Press, 1993.

Asger Kirkeby-Hinrup, Sascha Benjamin Fink, and Mads Overgaard. The multiple generator hypothesis. *Neuroscience of Consciousness*, 2025(1):niaf035, 2025.

Heinrich Klüver. *Mescal and Mechanisms of Hallucinations*. University of Chicago Press, 1966.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

Stephen LaBerge. *Lucid Dreaming*. Ballantine Books, 1985.

Victor A. F. Lamme. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11):494–501, 2006.

Victor A. F. Lamme. How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, 1(3):204–220, 2010.

Hakwan Lau and David Rosenthal. Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8):365–373, 2011.

Joseph Levine. Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(4):354–361, 1983.

Benjamin Libet. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8(4):529–539, 1985.

Rodolfo R. Llinás and Urs Ribary. Coherent 40-Hz oscillation characterizes dream state in humans. *Proceedings of the National Academy of Sciences*, 90(5):2078–2081, 1993.

Rodolfo R. Llinás, Urs Ribary, Diego Contreras, and Carlos Pedroarena. The neuronal basis for consciousness. *Philosophical Transactions of the Royal Society of London B*, 353(1377): 1841–1849, 1998.

Robert Long, Jeff Sebo, Patrick Butlin, Jonathan Birch, David Chalmers, et al. Taking AI welfare seriously, 2024.

Lucia Melloni et al. An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE*, 18(2):e0268577, 2023.

Thomas Metzinger. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press, 2003.

Thomas Metzinger. *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books, 2009.

Boris Milinkovic and Jaan Aru. Biological computationalism. *Neuroscience & Biobehavioral Reviews*, 181:106524, 2025.

Martin M. Monti et al. Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine*, 362(7):579–589, 2010.

Thomas Nagel. What is it like to be a bat? *Philosophical Review*, 83(4):435–450, 1974.

Nature Neuroscience Editors. Concerns about integrated information theory. *Nature Neuroscience*, 2025. Editorial / response regarding IIT empirical status.

Adrian M. Owen et al. Detecting awareness in the vegetative state. *Science*, 313(5792):1402, 2006.

Roger Penrose and Stuart Hameroff. Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation*, 40 (3–4):453–480, 1994.

Yair Pinto et al. Split brain: Divided perception but undivided consciousness. *Brain*, 140(5): 1231–1237, 2017.

Viola Priesemann et al. Neuronal avalanches differ from wakefulness to deep sleep — evidence from intracranial depth recordings in humans. *PLOS Computational Biology*, 9(3):e1002985, 2013.

Viola Priesemann et al. Spike avalanches in vivo suggest a driven, slightly subcritical brain state. *Frontiers in Systems Neuroscience*, 8:108, 2014.

Antje A. T. S. Reinders et al. One brain, two selves. *NeuroImage*, 20(4):2119–2125, 2003.

Antje A. T. S. Reinders et al. Cross-examining dissociative identity disorder: Neuroimaging and etiology on trial. *Neurocase*, 14(1):44–53, 2008.

Antti Revonsuo. Binding and the phenomenal unity of consciousness. *Consciousness and Cognition*, 8(2):173–185, 1999.

Eugenio Rodriguez et al. Perception's shadow: Long-distance synchronization of human brain activity. *Nature*, 397(6718):430–433, 1999.

David Rosenthal. *Consciousness and Mind*. Oxford University Press, 2005.

Michael Schartner et al. Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific Reports*, 7:46421, 2017.

Aaron Schurger, Jacobo D. Sitt, and Stanislas Dehaene. An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences*, 109(42):E2904–E2913, 2012.

Eric Schwitzgebel. AI and consciousness, 2025.

Anil Seth. *Being You: A New Science of Consciousness*. Dutton, 2021.

Wolf Singer and Charles M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, 18:555–586, 1995.

Galen Strawson. Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10–11):3–31, 2006.

Enzo Tagliazucchi et al. Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis. *Frontiers in Physiology*, 3:15, 2012.

Enzo Tagliazucchi et al. Increased global functional connectivity correlates with LSD-induced ego dissolution. *Current Biology*, 26(8):1043–1050, 2016.

Max Tegmark. Importance of quantum decoherence in brain processes. *Physical Review E*, 61(4):4194–4206, 2000.

Christopher Timmermann et al. Neural correlates of the DMT experience assessed with multivariate EEG. *Scientific Reports*, 9:16324, 2019.

Christopher Timmermann et al. Human brain effects of DMT assessed via EEG-fMRI. *Proceedings of the National Academy of Sciences*, 120(13):e2218949120, 2023.

Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5:42, 2004.

Giulio Tononi, Larissa Albantakis, Leonardo Barbosa, et al. Consciousness or pseudo-consciousness? A clash of two paradigms. *Nature Neuroscience*, 28:694–702, 2025.

Anne Treisman. The binding problem. *Current Opinion in Neurobiology*, 6(2):171–178, 1996.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

John von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, 1932.

Daniel M. Wegner. *The Illusion of Conscious Will*. MIT Press, 2002.

Eugene P. Wigner. Remarks on the mind-body question. In I. J. Good, editor, *The Scientist Speculates*. Heinemann, 1961.

Stephen Wolfram. *A New Kind of Science*. Wolfram Media, 2002.

Wojciech H. Zurek. Decoherence, einselection, and the quantum origins of the classical. *Reviews of Modern Physics*, 75(3):715–775, 2003.