Toward a Mathematical Formalization of the Four-Model Theory: A Recommended Approach

Matthias Gruber

Independent researcher

ORCID: 0009-0005-9697-1665

Correspondence: matthias@matthiasgruber.com

Abstract

The Four-Model Theory of Consciousness (FMT; Gruber, 2015, 2026) proposes that consciousness is constituted by real-time self-simulation across four nested models — Implicit World Model (IWM), Implicit Self Model (ISM), Explicit World Model (EWM), and Explicit Self Model (ESM) — operating at the edge of chaos. The theory currently operates in natural language. This paper outlines a recommended mathematical formalization strategy. Crucially, the four models are understood as a *minimum sufficient set* for human-level consciousness, not an exhaustive enumeration: the biological substrate — built from spiking neurons atop proteomic networks with their own intracellular learning — implements an uncountable number of overlapping models on both sides of the implicit/explicit divide. The formalization must therefore treat models not as discrete, countable objects but as a continuous density over a model space, with the virtual/non-virtual split as the one hard ontological boundary. Six formalization modules are proposed: (1) a continuous model-space framework replacing the discrete 2×2 taxonomy, (2) permeability as an information-theoretic quantity, (3) criticality operationalization, (4) ESM redirection dynamics, (5) self-referential closure, and (6) category-theoretic architecture. A phased build order prioritizes empirically testable components. The formalization is offered as a research program specification for mathematically trained collaborators; verification of the formal apparatus is explicitly deferred to domain experts.

Keywords: consciousness, formalization, four-model theory, model space, criticality, transfer entropy, self-referential closure, information geometry

---

1. Introduction

1.1 The Formalization Gap

The Four-Model Theory (Gruber, 2015, 2026) addresses all eight canonical requirements for a theory of consciousness — 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 — within a single framework built on five principles: criticality, virtual qualia, a redirectable Explicit Self Model, variable implicit-explicit permeability, and virtual model forking.

However, the theory currently operates entirely in natural language. This is its single largest vulnerability. Integrated Information Theory (IIT; Tononi, 2004; Albantakis et al., 2023) has Φ and the qualia space formalism. Predictive Processing (PP; Friston, 2010; Seth, 2021) has the free energy principle and active inference. Global Neuronal Workspace (GNW; Dehaene & Changeux,

2011) has computational broadcasting models. The Four-Model Theory has verbal descriptions of mechanisms — "permeability increases," "ESM redirects," "criticality threshold" — without specifying what these quantities *are* in measurable, falsifiable terms.

Without formalization, the theory's flexibility is a liability: it can potentially accommodate any observation post hoc, which undermines its empirical credentials. The theory needs mathematical precision to constrain its claims, generate quantitative predictions, and interface with the existing formal landscape of consciousness science.

1.2 Why Not Simply Formalize the Four Models?

The obvious approach — assigning a mathematical object to each of the four models and formalizing their interactions — would be premature and, importantly, would misrepresent the theory's own commitments.

The four models (IWM, ISM, EWM, ESM) are a *principled lower bound*, not an architectural specification. They represent the minimum configuration required for human-level consciousness: a system must model both the world and itself, and must do so at both the structural/implicit and the dynamic/explicit level. This is a constraint on the *minimum*, not a claim about the *actual*.

The biological substrate — spiking neurons atop proteomic networks, with intracellular signaling pathways constituting their own learning and computational intelligence even within a single cell (Bhatt et al., 2009; Bhalla, 2014) — implements an effectively uncountable number of overlapping models on both sides of the implicit/explicit divide. A motor model of reaching for a cup simultaneously encodes world-geometry (cup location, obstacle positions) and self-kinematics (arm configuration, grip aperture). An emotional model of a social interaction simultaneously encodes other-knowledge (world) and self-assessment (self). The boundaries between "models" are not sharp and their number is not well-defined in a neural architecture.

To formalize the theory correctly, therefore, requires treating the four models not as four discrete objects but as four *regions* in a continuous space of modeling activity — the canonical extrema around which the actual, high-dimensional, uncountable modeling ecology of the brain is organized. The formalization must be statistical, not enumerative.

1.3 Scope and Limitations

This paper specifies a formalization research program. It proposes mathematical frameworks, defines quantities, and outlines a build order. It does *not* verify the mathematical apparatus — the author is not a mathematician, and formal verification is explicitly deferred to domain experts. The equations presented here are intended as precise specifications of *what needs to be formalized* and *which mathematical tools are appropriate*, not as proven theorems.

The paper is structured as follows. Section 2 develops the continuous model-space framework. Section 3 formalizes permeability. Section 4 addresses criticality operationalization. Section 5 formalizes the ESM redirection mechanism. Section 6 addresses self-referential closure. Section 7 outlines a category-theoretic architecture. Section 8 proposes a phased build order. Section 9 identifies what formalization would buy and what it cannot deliver.

---

2. From Discrete Models to a Continuous Model Space

2.1 The Model Space

Instead of four enumerable models, define a **model space** M — a high-dimensional manifold where every point represents a distinct model the substrate is running. Each model m ∈ M has two primary properties:

- **Scope** s(m) ∈ [0, 1]: a continuous axis from pure self-representation (0) to pure world-representation (1).
- **Mode** ν(m) ∈ [0, 1]: a continuous axis from fully implicit/structural (0) to fully explicit/phenomenal (1).

The scope axis captures the observation that most actual neural models blend self and world: a motor reaching model encodes both world-geometry and body-kinematics simultaneously. The mode axis captures the theory's central claim that the implicit-explicit boundary is graded, not binary (Gruber, 2026, Section 3.6) — while maintaining that there exists a threshold ν_crit above which modeling activity is phenomenal.

The four canonical models are recovered as **extremal points** in this 2D projection:

- IWM ≈ (s = 1, $\nu = 0$): Pure world-knowledge, fully implicit.
- ISM ≈ (s = 0, $\nu = 0$): Pure self-knowledge, fully implicit.
- EWM ≈ (s = 1, $\nu = 1$): Pure world-simulation, fully explicit.
- ESM ≈ (s = 0, $\nu = 1$): Pure self-simulation, fully explicit.

But the actual substrate populates the *entire* $[0, 1]^2$ space with a density of models.

2.2 The Model Density Function

Define a **model density** ρ(s, ν, t) over the scope-mode space at time t. The quantity ρ(s, ν, t) ds dν represents the "amount of modeling activity" in the region [s, s+ds] × [ν, ν+dν] at time t. This is a measure-theoretic quantity, analogous to a probability density but not normalized to 1 — the total modeling activity of the brain can vary across states (more total modeling during waking than during deep sleep).

The virtual/non-virtual split — the one hard ontological claim of the theory — becomes a threshold on the mode axis:

- Virtual (phenomenal): $\nu > \nu$_crit
- Real (substrate): $\nu < \nu$_crit

Everything above ν_crit is part of the conscious simulation. Everything below it operates "in the dark." The threshold ν_crit is not a convention but an ontological boundary — the theory claims this boundary is real, even if its precise value must be determined empirically.

2.3 The Minimum-Configuration Constraint

The theory's claim about the four canonical models becomes a density constraint rather than a discrete architectural specification:

For consciousness of the human type, ρ must have significant mass near all four extremal points in (s, ν) space.

Formally, define four regions R_IWM, R_ISM, R_EWM, R_ESM as neighborhoods of the four corners, and require:

$\int_{R_k} \rho(s, \nu, t)\, ds\, d\nu > \theta_k$ for each k ∈ {IWM, ISM, EWM, ESM}

where θ_k are minimum-mass thresholds. A system with mass only near ($s = 1$, $\nu = 1$) — world-simulation without self-simulation — would not be conscious. A system with mass only below ν_{crit} — implicit processing without explicit simulation — would not be conscious. The four-model requirement is a constraint on the density profile, not a count of discrete objects.

This formulation accommodates the uncountable-models reality of the biological brain while preserving the theory's principled claim about the minimum sufficient configuration.

2.4 The Hierarchical Depth Axis

The model density function should be extended to account for the five-system hierarchy (Gruber, 2015, 2026):

$\rho(s, \nu, \ell, t)$

where $\ell \in \{1, 2, 3, 4, 5\}$ indexes the hierarchical level:

1. Physical system (atoms, molecules)
2. Electrochemical system (ion gradients, action potentials)
3. Proteomic system (receptor configurations, protein expression, intracellular signaling)
4. Topological system (synaptic connectivity, circuit architecture)
5. Virtual system (real-time dynamical patterns, the cortical automaton)

The proteomic layer ($\ell = 3$) is particularly relevant: receptor configurations encode prediction models about neurotransmitter dynamics, protein expression patterns constitute a slow, molecular-scale computational intelligence operating on timescales of minutes to days (Bhatt et al., 2009). These are implicit models in their own right, but at a level below the topological/synaptic level at which the standard four-model description operates.

The real/virtual split remains clean — it is a threshold on ν — but the implicit side has stratified depth, with models within models within models, all the way down to proteomic computation. The four-model description is the top-level statistical summary, analogous to describing a fluid by temperature and pressure when the underlying molecular dynamics are vastly more complex.

---

3. Permeability as an Information-Theoretic Quantity

3.1 Transfer Entropy Formulation

The implicit-explicit boundary and its variable permeability are arguably the most explanatorily productive mechanism in the theory, underlying the accounts of psychedelic phenomenology, anosognosia, dreams, meditation, and hypnagogia (Gruber, 2026, Section 3.6). Formalizing permeability is therefore the highest-priority task.

Define **permeability** P as the transfer entropy from implicit model parameters to explicit model states. Transfer entropy (Schreiber, 2000) measures directed information flow between time series:

$T_{X \to Y} = \Sigma\, p(y_{t+1}, y_t, x_t) \log [p(y_{t+1} \mid y_t, x_t) \,/\, p(y_{t+1} \mid y_t)]$

Applied to the model-space framework:

$P_{\text{implicit}\to\text{explicit}}(s, t) = T_{\{\rho(s, \nu<\nu_{\text{crit}}) \to \rho(s, \nu>\nu_{\text{crit}})\}}(t)$

This captures what the theory means by permeability: how much normally-implicit information (below ν_{crit}) is making it into the conscious simulation (above ν_{crit}) at scope position s and time t. Permeability can be global (averaged over all s) or local (at a specific scope position), which naturally maps the theory's distinction between global permeability changes (psychedelics) and local permeability deficits (anosognosia).

3.2 Permeability Profiles as Quantitative Predictions

The theory's phenomenological claims become quantitative predictions about the permeability profile:

State	Global P	Local P anomalies	Predicted measurement
Normal waking	Medium	None (selective gating)	Baseline transfer entropy
Psychedelic (psilocybin, LSD)	High	None (global increase)	Transfer entropy spike, broadband
Salvia divinorum (high dose)	High + ESM disruption	Self-scope P collapses	$T_{\{\text{ISM}\to\text{ESM}\}} \to 0$; $T_{\{\text{IWM}\to\text{ESM}\}} \uparrow$
Anosognosia (right hemisphere stroke)	Normal (most domains)	Low at damaged domain	Domain-specific transfer entropy deficit
Deep sleep (NREM)	Very low	Uniform suppression	Near-zero transfer entropy
REM dreaming	Medium	Internally driven (from W, not from s(t))	T from parameters high; T from sensory input ≈ 0
Meditation	Selectively elevated	Trained domain-specificity	Targeted transfer entropy increases
Pre-sleep / hypnagogia	Gradually increasing	Bottom-up (low-level $\to$ high-level)	Hierarchical onset of transfer entropy

These predictions are directly testable with existing neuroimaging data — transfer entropy can be estimated from EEG, MEG, and fMRI time series (Vicente et al., 2011; Wibral et al., 2014).

3.3 The Gating Operator

More formally, define a gating function g: $W \times X \to [0, 1]^N$ that modulates how much implicit content influences explicit dynamics. The substrate dynamics (Equation [1] of Section 4.1) become:

$x(t+1) = f(x(t), s(t), g(W, x(t)) \odot W)$

where $\odot$ is element-wise multiplication. The gating function g is the permeability operator. Psychedelics globally increase g toward 1; anosognosia locally decreases g in specific scope bands; general anesthesia (propofol) drives g toward 0.

The gating function g itself depends on: - The substrate state x(t): attentional gating (top-down control of what becomes conscious). - Neurotransmitter dynamics: pharmacological gating (serotonergic agonism increases g globally; GABAergic agonism decreases g globally). - Structural integrity: lesion-dependent gating (stroke damage eliminates g in specific pathways).

This connects naturally to the criticality-rhythm relationship that the theory identifies as an open question (Gruber, 2026, Section 9).

3.4 The Fokker-Planck Dynamics

The dynamics of the model density ρ under variable permeability can be modeled as a Fokker-Planck equation on the model space:

$\partial\rho/\partial t = -\nabla \cdot (v\ \rho) + D\ \nabla^2\rho + S(s, \nu, t)$

where: - $v(s, \nu, t)$ is a drift field: deterministic migration of models along the scope-mode axes. Attention directs drift along ν (making implicit content explicit); context shifts direct drift along s (shifting between self-focused and world-focused processing). - D is a diffusion coefficient: stochastic leakage across the implicit-explicit boundary — the baseline permeability noise that produces phenomena like visual snow and spontaneous phosphenes. - $S(s, \nu, t)$ is a source/sink term: creation and destruction of models. Learning creates new implicit models (increases ρ below ν_crit); forgetting destroys them; sensory input injects new explicit models (increases ρ above ν_crit).

State transitions then have specific signatures: - **Psychedelics**: Global increase in the drift velocity v_ν toward high ν. - **Propofol**: Collapse of D and v_ν to zero, with an absorbing boundary at ν_crit. - **Meditation**: Trained, selective control over $v(s, \nu, t)$ — the meditator learns to steer the drift field. - **Sleep onset**: Gradual reduction of v_ν combined with increasing D (controlled drift toward implicitness, with increasing stochastic permeability producing hypnagogic imagery).

3.5 Total Conscious Content

The total conscious content at time t — a single scalar representing the "amount" of phenomenal modeling — is:

$C(t) = \int_0^1 \int_{\{\nu\_crit\}}^1 \rho(s, \nu, t)\ d\nu\ ds$

This is directly analogous to what the Perturbational Complexity Index (PCI; Casali et al., 2013) and Lempel-Ziv complexity attempt to measure empirically. The formalization predicts that C(t) should correlate with PCI and similar complexity measures across consciousness states.

---

4. Criticality Operationalization

4.1 The Substrate as a Dynamical System

Define the full substrate state at time t as $x(t) \in X \subseteq R\hat{}N$, where N is the number of functional units (cortical columns in the biological brain). The dynamics follow:

$x(t+1) = f(x(t), s(t), W)$

where s(t) is sensory input and W is the connectivity matrix (synaptic weights). This is the cortical automaton — a discrete dynamical system on a high-dimensional lattice (Gruber, 2015, 2026).

The implicit models are the *parameters* of this system: - IWM = the world-knowledge partition of W. - ISM = the self-knowledge partition of W.

The explicit models are *patterns of activity* — projections of the state vector: - EWM(t) = Π_EWM · x(t) - ESM(t) = Π_ESM · x(t)

where Π_EWM and Π_ESM are projection operators. This maps the real/virtual split directly: W (parameters, slow, structural) = real side; x(t) projected through Π_EWM and Π_ESM (states, fast, transient) = virtual side.

4.2 Three Candidate Criticality Measures

The theory specifies Wolfram Class 4 / edge of chaos as the criticality requirement (Gruber, 2015, 2026). Class 4 is defined for cellular automata, not for continuous, noisy, high-dimensional neural substrates. Three existing measures can operationalize the requirement:

Branching ratio σ: The average number of descendant activations per ancestor activation across neuronal avalanches (Beggs & Plenz, 2003; Shew & Plenz, 2013). - $\sigma < 1$: subcritical (activity dies out) → Wolfram Class 1/2 - $\sigma = 1$: critical → Wolfram Class 4 - $\sigma > 1$: supercritical (activity explodes) → Wolfram Class 3

The ConCrit framework (Algom & Shriki, 2026) established that σ tracks consciousness across 140 datasets. The formalized claim: **Consciousness requires** $\sigma \in [\sigma\_\text{low}, \sigma\_\text{high}]$ where σ_low ≈ 0.95 and σ_high ≈ 1.1 (slightly subcritical to slightly supercritical, consistent with Priesemann et al., 2013, 2014).

Maximum Lyapunov exponent λ_max: For edge-of-chaos dynamics specifically (Bertschinger & Natschläger, 2004; Boedecker et al., 2012): - λ_max < 0: ordered (stable attractors) - λ_max ≈ 0: edge of chaos - λ_max > 0: chaotic

Detrended Fluctuation Analysis (DFA) exponent α: Long-range temporal correlations in neural time series (Hardstone et al., 2012). $\alpha \approx 1$ indicates critical dynamics with scale-free temporal structure.

4.3 Disambiguating Criticality Types

An important caveat: Kanders et al. (2017) demonstrated that avalanche criticality ($\sigma \approx 1$) and edge-of-chaos criticality (λ_max ≈ 0) do not necessarily co-occur in neural networks. These are measuring *different* phase transitions. The theory's reference to "edge of chaos" via Wolfram's Class 4 aligns more naturally with the Lyapunov exponent than with the branching ratio, while the empirical criticality literature (ConCrit) focuses primarily on the branching ratio and avalanche statistics.

The formalization must resolve this ambiguity. Three options:

1. **Both required**: Consciousness requires $\sigma \approx 1$ AND λ_max ≈ 0. This is the most restrictive and potentially the most empirically productive — it predicts that systems at avalanche criticality but not at edge-of-chaos (or vice versa) should not be conscious.
2. **Avalanche criticality sufficient**: The branching ratio is the operative measure; edge-of-chaos is a correlate but not independently required. This aligns with the ConCrit literature.

3. **Edge-of-chaos is primary**: λ_max ≈ 0 is the fundamental requirement (following Wolfram's framework most closely); avalanche criticality is a consequence. This aligns with the theoretical derivation in Gruber (2015).

Empirical resolution: compare PCI (or another consciousness-tracking measure) against σ and λ_max independently, particularly in states where the two measures diverge.

4.4 The Two-Threshold Formalization

The theory's two thresholds (criticality + self-simulation architecture) formalize as a conjunction:

Consciousness $\iff$ [$\sigma \in [\sigma$_low, σ_high]] $\wedge$ [$\exists$ significant ρ near all four corners of (s, ν) space above the minimum-mass thresholds]

Both conditions must hold simultaneously. A substrate at criticality but without self-modeling (e.g., a sandpile at the critical point) has complex dynamics but no consciousness. A system with the right architecture but below criticality (e.g., a brain under deep propofol anesthesia) has the structural capacity for consciousness but cannot instantiate the simulation.

---

5. ESM Redirection Dynamics

5.1 The ESM as an Input-Dependent Attractor

The theory's most distinctive prediction — that ego dissolution content is controllable via sensory input (Gruber, 2026, Prediction 3) — requires a formal account of how the ESM latches onto alternative inputs when normal self-referential input is disrupted.

Model the ESM-relevant region of the model density as a dynamical subsystem whose attractor landscape depends on its inputs. Let e(t) represent the aggregate ESM state (the integral of ρ over the self-scope, above-ν_crit region). Under normal conditions:

e(t+1) = h(e(t), i_self(t))

where i_self(t) is the normal self-referential input (interoceptive, proprioceptive). The ESM has a stable attractor basin around the "normal self" configuration.

During ego dissolution (high-dose psychedelic, salvia divinorum), i_self is disrupted:

e(t+1) = h(e(t), $\alpha \cdot$ i_self(t) + $(1 - \alpha) \cdot$ i_ext(t))

where $\alpha \to 0$ as dose increases and i_ext(t) is the dominant external input. The ESM's attractor basin reshapes around whatever i_ext dominates. This is the formal mechanism underlying the salvia phenomenology described in Gruber (2026, Section 6.1): users "become" objects in their environment because the ESM latches onto the strongest available sensory input.

5.2 Quantitative Predictions

The mutual information between the ESM state and its inputs provides the quantitative measure:

- I(e; i_self): How much the ESM state correlates with self-referential input.
- I(e; i_ext): How much the ESM state correlates with external sensory input.

The theory predicts: - Normal waking: I(e; i_self) » I(e; i_ext). The ESM is driven primarily by self-referential input. - Low-dose psychedelic: I(e; i_self) decreases; I(e; i_ext) begins to increase. - High-dose (ego dissolution): I(e; i_self) → 0; I(e; i_ext) → I_max. The ESM identity tracks the dominant external input. - Controllability prediction: In a controlled sensory environment during ego dissolution, I(e; i_ext) should track experimentally manipulated sensory input — visual, auditory, or proprioceptive — in a predictable, dose-dependent manner.

This is testable in principle with fMRI: measure the correlation between default mode network activity (proxy for ESM) and controlled sensory input versus interoceptive input across dose levels.

5.3 The Density Migration Account

In the model-space framework, ESM redirection corresponds to a specific density migration: ρ at the self-scope (s ≈ 0), explicit ($\nu > \nu$_crit) corner migrates toward the world-scope (s → 1) region while remaining above ν_crit. The system still runs an explicit simulation — consciousness is preserved — but the simulation's content shifts from self-dominated to world-dominated. Ego dissolution is not ESM abolition but ESM re-sourcing.

This density migration should be measurable as a shift in representational content within the networks that sustain explicit processing (default mode network shifting toward content typically associated with sensory processing networks), detectable via representational similarity analysis (RSA) of fMRI data during controlled psychedelic administration.

---

6. Self-Referential Closure

6.1 The Load-Bearing Argument

The theory's central philosophical claim — that self-referential closure is what makes the simulation phenomenal — carries the weight of the Hard Problem dissolution (Gruber, 2026, Section 3.4). A weather simulation models weather but does not model itself modeling weather; therefore it has an "outside" from which it can be fully described. A self-referential simulation at criticality has no such outside — the simulation *is* its own observer, and observation-from-inside is what the theory identifies as experience.

This argument needs formal grounding. Three mathematical approaches are proposed:

6.2 Self-Referential Depth via Recursive Representation

Define a representation map ρ_n that encodes how a subsystem models another at recursion depth n:

ρ_n: X_ESM → M(X_ESMˆ(n-1))

where M denotes "models of" and the superscript denotes recursion level. The graduated levels of consciousness (Gruber, 2015, 2026) then formalize as:

Level	Recursion depth	Description
Basic consciousness	ρ_0	ESM represents the EWM
Simply extended	ρ_1	ESM represents itself representing the EWM

Level	Recursion depth	Description
Doubly extended	ρ_2	ESM represents itself representing itself representing the EWM
Triply extended	ρ_3	Third-order recursion; Meta-Problem arises here

The **self-knowledge measure** quantifies how well the system predicts its own next state:

R = 1 − H(e(t+1) | ê(t+1)) / H(e(t+1))

where ê(t+1) is the system's *own prediction* of its next ESM state. R = 1 means perfect self-prediction; R = 0 means no self-knowledge.

6.3 Self-Referential Closure as a Fixed Point

The most compact formalization: consciousness occurs when the ESM reaches a **fixed point** of self-representation. Define a map Φ: M → M where Φ(m) is "the model of m." Self-referential closure occurs when:

Φ(m) = *m*

The self-model models itself — the model and the modeled coincide. This is a well-defined mathematical object (a fixed point of a self-referential map), related to Lawvere's fixed-point theorem (Lawvere, 1969) and Kauffman's self-referential forms (Kauffman, 2005).

The theory's claim that "the simulation *is* the thing being simulated" (Gruber, 2026, Section 3.4) is precisely this fixed-point condition. At a fixed point of self-representation, there is no remainder — no "outside view" from which the process can be fully described without participating in it.

The critical connection to the Hard Problem dissolution: the category error identified by the theory (seeking phenomenality at the substrate level) becomes formally precise at the fixed-point level. The phenomenal properties exist at m*, not at the substrate level that computes* Φ*. The function* Φ *runs on the substrate; the fixed point m* is a property of the dynamics, not of any particular substrate element.

6.4 Complexity Cost of Self-Reference

Self-referential modeling has a computational overhead. For a system to achieve recursion depth n, its total complexity must satisfy:

C_total ≥ Σ_{k=0}^{n} C_k + overhead(n)

where C_k is the complexity required for level-k representation and the overhead grows with depth. This explains why triply-extended consciousness requires a large, complex substrate (human-scale cortex) while simpler organisms support only basic consciousness — they lack the computational overhead for deeper recursion. It also provides a formal account of why the six-layer neocortex exceeds the three-layer minimum for universal function approximation (Cybenko, 1989): the additional layers provide the overhead needed for recursive self-simulation (Gruber, 2015).

7. Category-Theoretic Architecture

7.1 Two Categories

Category theory provides the most natural language for the theory's architectural claims because it is designed to formalize structural relationships between mathematical objects (Mac Lane, 1998; Tsuchiya, Taguchi, & Saigo, 2016).

Define two categories:

- **Sub** (Substrate): Objects are substrate states (W, x(t)). Morphisms are physical dynamics — the state transitions described by the dynamical system equations.
- **Sim** (Simulation): Objects are virtual-model states (EWM(t), ESM(t), or equivalently, the above-ν_crit portion of ρ). Morphisms are experiential transitions — phenomenal changes.

7.2 The Consciousness Functor

Consciousness is a **functor** F: Sub → Sim that maps substrate dynamics to simulation dynamics while preserving compositional structure:

- F maps physical state transitions to experiential transitions.
- F preserves composition: if substrate state A transitions to B and B to C, the experiential transition A→C is the composition of the experiential transitions A→B and B→C.

The real/virtual split is the distinction between the domain category (Sub) and the codomain category (Sim). The Hard Problem dissolution becomes: seeking phenomenal properties in Sub is a category error because phenomenal properties exist in Sim. The functor F *generates* Sim from Sub, but Sim has properties (qualia, unity, selfhood) that do not exist as properties of Sub — just as the image of a functor can have properties not present in the domain.

7.3 Permeability as a Natural Transformation

Variable permeability can be formalized as a **natural transformation** η: F_normal ⇒ F_altered between different consciousness functors. Under psychedelics, the functor changes — more substrate structure maps into the simulation — but the structural relationships are preserved (content appears in hierarchical order, V1 → V2/V3 → higher areas, not as random noise). The natural transformation ensures this structural preservation.

Smithe's (2024) structured active inference framework, which formalizes active inference using categorical systems theory and treats interfaces as compositional abstractions of Markov blankets, provides a directly applicable technical foundation. The FMT's implicit-explicit boundary could be modeled as a structured interface in this sense — a compositional boundary through which information flows, with the gating function g specifying the interface's permeability properties.

7.4 Forking as Coproduct

The theory's virtual model forking — the mechanism underlying dissociative identity disorder (Gruber, 2026, Section 6.2) — formalizes as a **coproduct** in the Sim category. A single substrate object in Sub maps to multiple simulation objects in Sim:

$F(x(t)) = ESM\_1(t) \sqcup ESM\_2(t) \sqcup \ldots \sqcup ESM\_n(t)$

where $\sqcup$ is the coproduct (disjoint union). This captures the claim that DID involves a single substrate running multiple ESM configurations, each constituting a distinct experiential self, with only one active at any given time.

---

8. Phased Build Order

The formalization project is substantial. A pragmatic build sequence, ordered by empirical accessibility and mathematical difficulty:

Phase 1: Highest Priority (Directly Testable with Existing Data)

Module 3.2 — Permeability as transfer entropy: Compute transfer entropy between neural signals in known "implicit" processing regions and known "explicit"/"conscious access" regions across existing neuroimaging datasets from psychedelic, anesthesia, sleep, and meditation studies. Test the permeability profile predictions (Table 1) quantitatively. This requires no new mathematical development — transfer entropy estimation is well-established (Wibral et al., 2014).

Module 4 — Criticality threshold mapping: The branching ratio σ and Lyapunov exponent λ_max are already measured in the ConCrit literature (Algom & Shriki, 2026; Hengen & Shew, 2025). Map FMT's predictions onto existing datasets. Attempt to disambiguate avalanche criticality from edge-of-chaos criticality in consciousness contexts (Section 4.3).

Phase 2: Core Formalism (Requires Dedicated Mathematical Work)

Module 2 — Continuous model space: Define the model density function ρ(s, ν, t) rigorously. Specify how to estimate ρ from neuroimaging data using representational similarity analysis (RSA), encoding models, and dimensionality reduction techniques (ICA, NMF). Build a minimal computational model (recurrent spiking network) and estimate ρ from its activity.

Module 5 — ESM redirection: Formalize the attractor-switching mechanism. Build a minimal computational model with a self-model unit and demonstrate input-dependent identity switching under perturbation. Generate quantitative predictions for the salvia divinorum controlled-input experiment (Gruber, 2026, Prediction 3).

Phase 3: Deep Formalism (Hardest, Highest Potential Impact)

Module 6 — Self-referential closure: The fixed-point formalization. This requires working at the intersection of dynamical systems theory and mathematical logic. The connection to Lawvere's fixed-point theorem needs rigorous development. Investigate whether the fixed-point condition can be shown to formally entail properties that correspond to the "no outside view" argument.

Module 7 — Category-theoretic architecture: The functor construction. This is the most elegant but also the most abstract. Collaboration with a category theorist who has consciousness theory exposure (cf. Tsuchiya et al., 2016; Smithe, 2024) is recommended.

Phase 4: Computational Validation

Minimal FMT simulator: A recurrent spiking neural network with: - Tunable criticality (via connectivity scaling to $\sigma \approx 1$) - A measurable model density ρ(s, ν, t) estimated via ICA/NMF - A

gating function g implementing variable permeability - A self-referential loop (output fed back as input through a self-model pathway)

Demonstrate the key phenomena in silico: ESM redirection under self-input disruption, permeability-dependent content flooding, criticality-dependent simulation coherence, holographic degradation under network bisection. This does not prove consciousness — but it demonstrates that the formalized theory's dynamics behave as predicted.

---

9. What Formalization Buys — And What It Cannot

9.1 What It Buys

Constraint: Verbal descriptions are flexible enough to accommodate post-hoc explanations. The Fokker-Planck dynamics, the transfer entropy measures, and the attractor-switching model commit to specific functional forms that can be empirically wrong.

Quantitative predictions: "Permeability increases under psychedelics" becomes "Transfer entropy from implicit to explicit regions increases by factor k at dose d" — a claim falsifiable with a number.

Simulation: A formalized theory can be implemented in code. Predictions can be tested computationally before committing to expensive neuroimaging experiments.

Interoperability: The transfer entropy measure connects to Predictive Processing's information-theoretic tools. The branching ratio connects to ConCrit. The category-theoretic framework connects to Smithe's structured active inference. Formalization turns FMT from an isolated framework into something interoperable with the rest of the field.

Sharpened dissolution: The fixed-point formalization either works or it doesn't. If self-referential closure can be made rigorous — if the fixed-point condition can be formally shown to entail properties corresponding to inside/outside asymmetry — that's a genuine philosophical contribution, not just a metaphor.

9.2 What It Cannot

Even the best formalization cannot derive phenomenality from mathematics. No equation will make someone who doubts consciousness understand what redness feels like. The value of formalization is not to solve the Hard Problem through mathematics but to make the theory's claims precise enough to be *clearly right or clearly wrong* — which is, in the end, the only honest standard a theory can meet.

The model-space approach introduces a further honest limitation: the model density $\rho(s, \nu, t)$ is a statistical description of something that may not be cleanly decomposable. The brain may not implement "models" in any separable sense — the activity patterns we decompose via ICA or NMF may be artifacts of our decomposition method rather than natural kinds. The formalization should therefore be understood as a measurement framework that makes the theory testable, not as a claim that the brain literally implements a model density function. The map is not the territory — but a good map is better than no map.

---

10. Conclusion

The Four-Model Theory requires mathematical formalization to constrain its claims, generate quantitative predictions, and interface with the existing formal landscape of consciousness science. The correct formalization strategy must respect the theory's own commitment: the four canonical models are a minimum sufficient set, not an exhaustive enumeration. The biological substrate implements an uncountable ecology of models, and the formalization must be statistical rather than enumerative.

The continuous model-space framework — with scope and mode as continuous axes, the virtual/non-virtual split as a threshold on the mode axis, and the Fokker-Planck equation governing the density dynamics — provides the necessary mathematical language. Combined with transfer entropy for permeability, established criticality measures, attractor dynamics for ESM redirection, and fixed-point theory for self-referential closure, the framework generates quantitative predictions that are testable with existing neuroimaging methods and computational models.

The formalization project is substantial but modular. Phase 1 (transfer entropy estimation and criticality mapping) can proceed immediately with existing tools and data. Phases 2–3 require dedicated mathematical collaboration. Phase 4 (computational validation) provides the ultimate demonstration that the formalized theory's dynamics behave as predicted.

This paper specifies the program. Its execution requires mathematically trained collaborators — and the intellectual honesty to discover that some of these formalizations may reveal the theory to be wrong in specific, identifiable ways. That would be a feature, not a bug.

References

Albantakis, L., et al. (2023). Integrated information theory (IIT) 4.0. *PLOS Computational Biology*, 19(10), e1011465.

Algom, S., & Shriki, O. (2026). The ConCrit framework: Critical brain dynamics as a unifying mechanism for consciousness theories. *Neuroscience & Biobehavioral Reviews*.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Beggs, J. M., & Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *Journal of Neuroscience*, 23(35), 11167–11177.

Bertschinger, N., & Natschläger, T. (2004). Real-time computation at the edge of chaos in recurrent neural networks. *Neural Computation*, 16(7), 1413–1436.

Bhalla, U. S. (2014). Molecular computation in neurons: a modeling perspective. *Current Opinion in Neurobiology*, 25, 31–37.

Bhatt, D. H., Zhang, S., & Bhatt, W. B. (2009). Dendritic spine dynamics. *Annual Review of Physiology*, 71, 261–282.

Boedecker, J., Obst, O., Lizier, J. T., Mayer, N. M., & Asada, M. (2012). Information processing in echo state networks at the edge of chaos. *Theory in Biosciences*, 131(3), 205–213.

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

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

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.

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

Gruber, M. (2015). *Die Emergenz des Bewusstseins*. Self-published.

Gruber, M. (2026). The Four-Model Theory of Consciousness: A Simulation-Based Framework Unifying the Hard Problem, Binding, and Altered States. *Zenodo* preprint. https://doi.org/10.5281/zenodo.18669891

Hardstone, R., et al. (2012). Detrended fluctuation analysis: a scale-free view on neuronal oscillations. *Frontiers in Physiology*, 3, 450.

Hengen, K. B., & Shew, W. L. (2025). Meta-analysis of neural criticality across 140 datasets. [Consolidated in ConCrit framework.]

Kanders, K., Lorimer, T., & Stoop, R. (2017). Avalanche and edge-of-chaos criticality do not necessarily co-occur in neural networks. *Chaos*, 27(4), 047408.

Kauffman, L. H. (2005). Self-reference and recursive forms. *Journal of Social and Biological Structures*, 10(1), 53–72.

Lawvere, F. W. (1969). Diagonal arguments and Cartesian closed categories. *Lecture Notes in Mathematics*, 92, 134–145.

Mac Lane, S. (1998). *Categories for the Working Mathematician* (2nd ed.). Springer.

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

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

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2), 461–464.

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Dutton.

Shew, W. L., & Plenz, D. (2013). The functional benefits of criticality in the cortex. *The Neuroscientist*, 19(1), 88–100.

Smithe, T. S. C. (2024). Structured active inference. Extended abstract.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5, 42.

Tsuchiya, N., Taguchi, S., & Saigo, H. (2016). Using category theory to assess the relationship between consciousness and integrated information theory. *Neuroscience Research*, 107, 1–7.

Vicente, R., Wibral, M., Lindner, M., & Pipa, G. (2011). Transfer entropy — a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1), 45–67.

Wibral, M., et al. (2014). *Directed Information Measures in Neuroscience*. Springer.

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.