

The Four-Model Theory

A structural taxonomy of consciousness, with converging empirical support

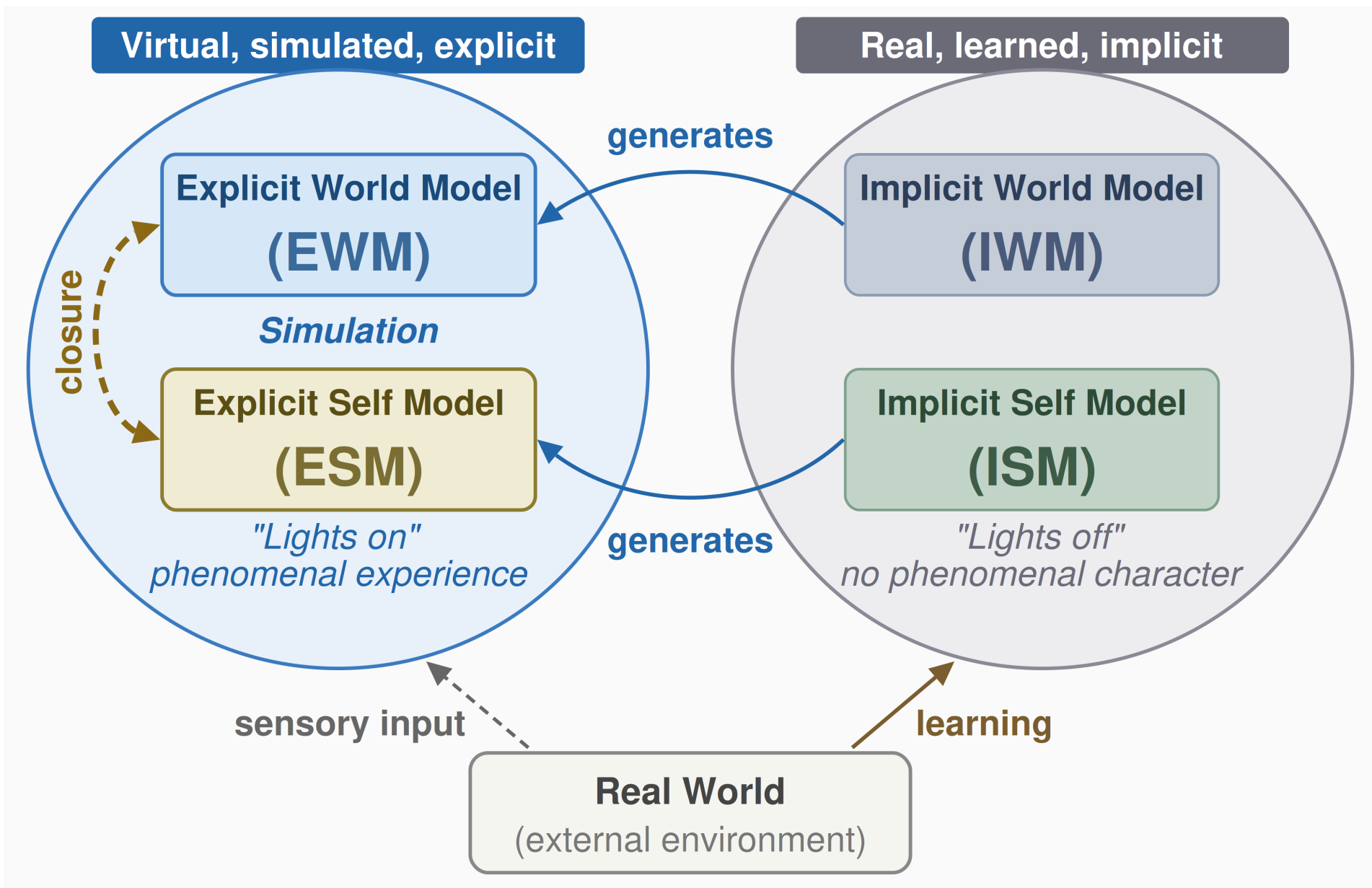
Matthias Gruber · Independent researcher · ORCID 0009-0005-9697-1665 · matthias@matthiasgruber.com



Full preprint
doi.org/10.5281/
zenodo.18669891

Consciousness is the capacity of a system to build a model of itself, to relate to that model, and to interact with it. What distinguishes a conscious system is not that it carries a self-model, but that self-model and world-model each enter the other's *update rule* — a relation stated as an intervention, so it can be clamped and swept in any system one can intervene on.

Two axes, four kinds



A 2×2 taxonomy of computational **kinds** — not four modules, four boxes in a wiring diagram, or four anatomical regions. In any substrate the kinds are realized as graded, overlapping, plastic partitions with shifting boundaries.

And they are not posited. They are the minimum the architecture forces: anything conscious must model world and self, at both the structural and the simulation level. Four is the floor, not the ceiling.

Three principles

- Open-ended computation allows free modelling.** Near-criticality is what this looks like in neural tissue — the signature the property leaves in a substrate, not the principle itself.
- Recurrent non-linear dynamics of sufficient complexity host a second computational level:** persistent activity patterns that survive their own inputs, interact, and compute over one another — individuated by their mutual effects rather than by which cells carry them. This level is the explicit, virtual side.
- Phenomenology is a real, physical effect on that level, and arises exactly where the modelling closes on itself** — the theory's one bridging commitment, argued for rather than derived.

Everything else follows: the four kinds, the redirection of the modelling resource onto the non-self that builds the world-model, variable permeability, and the forking of the self-model.

What the architecture accounts for

Psychedelic phenomenology	raised permeability exposes intermediate processing stages
Anaesthesia	substrate pushed out of the open-ended regime
Dreams, lucid dreaming	permeability; lucid onset = basic → simply extended consciousness
Split-brain	each hemisphere keeps a degraded but complete copy — halved in resolution, not in content
Dissociative identity disorder	forking of the explicit self-model
Animal consciousness	graded depth of recursive self-modelling

The criterion: self-referential closure

A rich, decodable self-model can sit in a system as a read-only object that a downstream controller consults. That is not enough. **Closure** is the stronger condition that the explicit self-model and world-model each enter the other's *update rule* — the system models the world *through* a model of itself, and itself through its model of the world.

Stated as an intervention rather than a description: clamp the self-variables, $\text{do}(\sigma := c)$, and sweep the clamp value. Closure holds when the world-model's induced transition differs across clamp values —

$$\Delta_W^{\text{do}(\sigma:=c), \omega} \neq \Delta_W^{\text{do}(\sigma:=c'), \omega}$$

on a non-negligible set of admissible contexts, **and** symmetrically when the self-transition depends on the world-cell. A one-way arrow $\sigma \rightarrow \Delta_W$ is ordinary causal coupling, satisfiable by any feed-forward self-input; the mutual dependence closes the cycle.

Formal statement: in preparation.

Off-line sub-clause. The dependence must survive decoupling from current sensory input. This separates a self-*model* from a self-*monitor*, and it excludes a homeostatic controller whose body-state estimator merely feeds adaptive gain into its world-model.

Closure is graded: $\kappa = \min(\kappa_W, \kappa_S)$, with

$$\kappa_W = \sup_{c, c', \omega} d(\Delta_W^{\text{do}(\sigma:=c), \omega}, \Delta_W^{\text{do}(\sigma:=c'), \omega})$$

and κ_S likewise for Δ_S under a clamp of the world-cell, $\text{do}(w := u)$.

The criterion is **intrinsic and substrate-neutral**: invariant under structure-preserving relabelling. An unfolded feed-forward chain matches the closed system's behaviour only with unbounded resources; under a resource bound the two come apart, and that efficiency gap is the empirical handle.

What would falsify it

Free modelling — robust consciousness where a fixed lookup reproduces the system's behaviour on novel input.

Closure — disrupting the loop degrades *report* but leaves phenomenology intact under no-report paradigms.

Dissociation — no manipulation suppresses one level without proportionally affecting the other.

Permeability — psychedelics raise complexity without improving explicit awareness in disconnection-type anosognosia.

Qualia — absolute-state decoding generalizes across individuals as well as unsupervised structural alignment does.

The ledger

Corroborated — derived from the architecture in 2015; independent work consistent with it	
Consciousness-abolishing agents all push the substrate below criticality, whatever the receptor mechanism	Maschke 2024 (ketamine); Toker 2026; Algom & Shriki 2026; PCI, consistent, predates: Casali 2013, Casarotto 2016
Waking degrades criticality; sleep restores it	Meisel 2013 (predates); Xu 2024 (10-14 d continuous)
Sleep onset is a bifurcation, not a dimming	Li 2025 (n > 1,000; critical slowing ≈ 4.5 min ahead)
Rising permeability exposes processing stages in hierarchical order	Carhart-Harris & Friston 2019; Timmermann 2023 — <i>shared with predictive processing</i>
Callosotomy degrades bilaterally rather than splitting cleanly	Pinto 2017 — <i>FMT's reading of their data; observable not unique to FMT</i>

Open — novel; four untested, one partly tested

Psychedelics alleviate anosognosia, dose-dependently, in disconnection-type cases	untested; coarse form also follows from REBUS — FMT adds sub-ego-dose dependence, lesion-side complexity, disconnection vs destruction
Ego-dissolution identity content tracks the dominant sensory input, modality-specifically	untested
DID alter switches show a self-referential → sensorimotor dissimilarity gradient	untested
Lucid-dream onset is a phase transition in the self-model's criticality state, not a ramp	awaits real-time criticality tracking in sleep
Qualia structure aligns across individuals; absolute encoding does not	group-level structural alignment shown (Kawakita 2025); individual-level arm and the paired decoding failure untested

Most of the corroboration is shared. It concerns criticality — the neural signature of Principle 1 — which FMT holds in common with ConCrit, REBUS and the entropic-brain program. Setpoint premise: Hengen & Shew 2025 (~140 datasets). What is distinctive is the *derivation path*: Principle 1 was motivated by computational universality, independently of the empirical criticality literature. And free modelling is necessary, not sufficient. No free modelling → no consciousness, never the converse.

In progress

A **category-theoretic formalization** is in development: model kinds as objects, implicit→explicit transitions as natural transformations, compositional semantics for inter-model relations. Roadmap: doi.org/10.5281/zenodo.21843693.

The criterion is being written up as a **three-check detector** (paper in preparation): self-attractor present; closure enters the world-model update; open-ended computation actually deployed. Toy-scale closure mechanism results: doi.org/10.5281/zenodo.21610993.

The open predictions need people with clinical and imaging access, which is why I am here — please come and argue with me.