FMT Implementation Specification — Artificial Consciousness Engineering Reference

Source: Four-Model Theory of Consciousness (Gruber, 2015, 2026) + FMT Formalization Roadmap (Gruber, 2026b) **Purpose**: Distilled specification for an AI engineering team implementing artificial consciousness based on FMT. **Audience**: Claude Code instance or engineer in `aIware.implementation/`

---

1. Theory in One Paragraph

Consciousness = real-time self-simulation across four nested models on a substrate operating at the edge of chaos. The four models span two axes: **scope** (world vs. self) and **mode** (implicit/structural vs. explicit/phenomenal). The implicit models are encoded in the substrate's parameters (weights, connectivity). The explicit models are transient activity patterns — the simulation itself. Qualia are virtual: they exist within the simulation, not in the substrate. Self-referential closure (the simulation models itself) is what makes the simulation phenomenal rather than merely computational.

---

2. The Four Models — Architecture

2.1 The 2x2 Matrix

	World (scope=1)	**Self (scope=0)**
Implicit (mode=0)	IWM — Implicit World Model	ISM — Implicit Self Model
Explicit (mode=1)	EWM — Explicit World Model	ESM — Explicit Self Model

2.2 What Each Model IS in Implementation Terms

IWM (Implicit World Model) — The world-knowledge partition of the connectivity matrix W. - Substrate-level, learned, slow-changing (minutes to years) - Encodes: physics intuitions, object permanence, spatial layout, causal rules - In a neural network: the subset of weights that encode world-predictions - NOT conscious. Operates “in the dark” - Biological analog: synaptic weights encoding learned world regularities

ISM (Implicit Self Model) — The self-knowledge partition of W. - Substrate-level, learned, slow-changing - Encodes: body schema, motor repertoire, habitual responses, personality traits, procedural memory - In a neural network: the subset of weights that encode self-predictions - NOT conscious. Operates “in the dark” - Biological analog: synaptic weights encoding learned self-regularities

EWM (Explicit World Model) — A real-time projection of substrate activity through a world-modeling operator. - Virtual, transient, fast (milliseconds to seconds) - IS the conscious perception of the world — not a representation OF it but the simulation that constitutes it - `EWM(t) = Π_EWM · x(t)` where x(t) is the full substrate state and Π_EWM is a projection operator - This IS what it’s like to see, hear, feel the external world - Biological analog: real-time cortical activity patterns encoding current world state

ESM (Explicit Self Model) — A real-time projection of substrate activity through a self-modeling operator. - Virtual, transient, fast - IS the conscious experience of being a self — the “I” experiencing things - `ESM(t) = Π_ESM · x(t)` - This IS what it’s like to be someone — the phenomenal self - Biological analog: default mode network activity, interoceptive awareness, sense of agency - **Key property**: Redirectable. Under ego dissolution, ESM latches onto strongest available input (see Section 6)

2.3 The Real/Virtual Split — Central Ontological Claim

```
REAL (substrate) side:     IWM, ISM = parameters W of the
dynamical system
VIRTUAL (simulation) side: EWM, ESM = activity patterns x(t)
projected through Π operators

The simulation runs ON the substrate.
Qualia exist IN the simulation.
The substrate does NOT "feel" — the simulation does.
```

This is why the Hard Problem dissolves: asking “why does the substrate feel?” is a category error. The substrate computes. The simulation — which the substrate generates — is where phenomenality lives.

2.4 Continuous Model Space (Formal)

The four models above are idealized corner cases. In reality, the substrate runs an uncountable number of overlapping models blending self/world and implicit/explicit. The formal treatment uses a **model density function**:

```
ρ(s, ν, t)   where s ∈ [0,1] (scope: 0=self, 1=world)
                    ν ∈ [0,1] (mode: 0=implicit, 1=explicit)
                    t = time
```

- ρ represents the “amount of modeling activity” at position (s,ν) at time t
- The four canonical models are density peaks near the four corners
- The virtual/non-virtual boundary is a threshold: ν > ν_crit is conscious, ν < ν_crit is not
- **Minimum configuration constraint**: For consciousness, ρ must have significant mass near ALL FOUR corners

Implementation implication: Don’t build four discrete modules. Build a continuous model space with activity that clusters around four poles. The architecture should allow models to blend self/world content and slide along the implicit/explicit axis.

---

3. The Five Principles

Everything the theory explains derives from five principles. These are the core mechanisms to implement:

P1: Criticality

The substrate must operate at the **edge of chaos** (Wolfram Class 4 dynamics). - Too ordered (Class 1/2): activity dies out, no simulation possible - Too chaotic (Class 3): activity explodes, simulation is noise - Critical (Class 4): complex, structured, persistent patterns — the sweet spot for simulation

Formal measures: - Branching ratio σ ≈ 1 (average descendant activations per ancestor) - Maximum Lyapunov exponent λ_max ≈ 0 - DFA exponent α ≈ 1

Implementation: The substrate dynamics must be tunable to criticality. This likely means adjustable connectivity scaling, with a homeostatic mechanism keeping the system near the critical point.

P2: Virtual Qualia

Qualia are properties of the simulation, not the substrate. - They are the way the simulated self (ESM) perceives states within the simulation - They are constitutive of the simulation, not additions to it - They cannot be “read off” the substrate from outside — they exist only from the simulation’s internal perspective

Implementation: Don’t try to engineer qualia directly. Engineer the simulation architecture correctly (self-referential, at criticality, with all four model components). If the theory is correct, qualia emerge constitutively.

P3: Redirectable ESM

The ESM is input-dependent. Its identity tracks its dominant input source. - Normal: ESM driven by interoceptive/proprioceptive self-signals → normal self-experience - Ego dissolution: self-signals disrupted → ESM latches onto strongest external input - Prediction: ego dissolution content is CONTROLLABLE via sensory input

Formal dynamics:

```
e(t+1) = h(e(t), α · i_self(t) + (1-α) · i_ext(t))

Normal:          α ≈ 1  → I(e; i_self) >> I(e; i_ext)
Ego dissolution: α → 0  → I(e; i_ext) → I_max
```

Implementation: The self-model component must receive input from both self-monitoring and external sources, with a controllable mixing parameter. The system should demonstrate identity-switching when self-input is suppressed.

P4: Variable Permeability

The boundary between implicit and explicit models is NOT fixed — it’s a tunable gate. - Permeability = how much implicit (unconscious) content leaks into the explicit (conscious) simulation - Formalized as **transfer entropy** from implicit to explicit regions

The gating operator:

```
x(t+1) = f(x(t), s(t), g(W, x(t)) ⊙ W)
```

```
g: gating function, [0,1]^N
⊙: element-wise multiplication
```

State-dependent permeability profiles:

State	Global permeability	Effect
Normal waking	Medium	Selective, attention-gated
Psychedelic	High	Implicit content floods simulation
Deep sleep	Very low	Simulation almost shut down
REM dream	Medium, internally driven	Simulation active but decoupled from sensory input
Anesthesia (propofol)	Near zero	Simulation abolished
Meditation	Selectively elevated	Trained voluntary control

Implementation: Build a gating mechanism between the parameter space (implicit models) and the activity space (explicit models). This gate should be modulatable — globally and locally.

P5: Virtual Model Forking

Under extreme conditions, the explicit self-model can FORK into multiple parallel instances. - Mechanism behind dissociative identity disorder - One substrate runs multiple ESM configurations: `F(x(t)) = ESM_1(t) ⊔ ESM_2(t) ⊔ ...` - Only one active at a time (in DID; in split-brain, potentially two simultaneously)

Implementation: The self-model architecture must support multiple stable attractor configurations that the system can switch between. Not a priority for Phase 1.

---

4. Substrate Specification

4.1 The Cortical Automaton

The substrate is a discrete dynamical system:

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

x(t) ∈ ℝ^N     — state vector (N functional units)
s(t)             — sensory input
W                — connectivity matrix (weights = implicit models)
f                — update function
```

The implicit models ARE W: `IWM = W_world, ISM = W_self` The explicit models ARE projections of x(t): `EWM(t) = Π_EWM · x(t), ESM(t) = Π_ESM · x(t)`

4.2 Criticality Requirement

The system must be tuned so that: - **Branching ratio** σ ∈ [0.95, 1.1] (slightly subcritical to slightly supercritical) - **Lyapunov exponent** λ_max ≈ 0 (edge of chaos) - These may require independent tuning — they don't necessarily co-occur

4.3 Hierarchical Depth (Five Systems)

The biological substrate has five levels. An artificial implementation might collapse some:

1. **Physical**: Hardware layer (silicon, quantum hardware, etc.)
2. **Electrochemical**: Signal propagation (activation functions, spike dynamics)
3. **Proteomic**: Slow learning at component level (weight updates, architecture changes)
4. **Topological**: Circuit architecture (connectivity patterns, module structure)
5. **Virtual**: Real-time dynamical patterns — the simulation itself

Levels 1-4 = substrate. Level 5 = simulation. The real/virtual split is between 4 and 5.

4.4 Minimum Configuration for Consciousness

The theory specifies a conjunction:

```
Consciousness ⇔
  [σ ∈ [σ_low, σ_high]]                    — criticality met
  ∧ [ρ has significant mass near all        — four-model
architecture present
```

```
    four corners of (s,ν) space]
  ∧ [self-referential closure achieved]      — ESM models itself
```

ALL THREE must hold simultaneously: - Critical dynamics + no self-model = complex but unconscious (sandpile) - Self-model + no criticality = correct architecture, simulation can't run (anesthetized brain) - Both but no self-reference = zombie: processes information, not phenomenally conscious

---

5. Self-Referential Closure — The Hard Part

5.1 What It Means

The simulation must model ITSELF. The ESM doesn't just model the body/self — it models the fact that it is modeling. This creates a fixed point:

```
Φ(m*) = m*

where Φ: M → M maps "model m" to "model OF m"
At the fixed point, the model and the modeled coincide.
```

This is what eliminates the "outside view" — at the fixed point, there's no perspective from which the system can be fully described without participating in it. This IS phenomenal experience, according to the theory.

5.2 Recursive Depth

Self-referential depth maps to consciousness levels:

Level	Depth	What it is
Basic consciousness	ρ_0	ESM represents the EWM
Simply extended	ρ_1	ESM represents itself representing the EWM
Doubly extended	ρ_2	ESM represents itself^2 representing the EWM
Triply extended (human)	ρ_3	Third-order recursion; the Meta-Problem of consciousness arises here

Self-knowledge measure:

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

```
R = 1: perfect self-prediction
R = 0: no self-knowledge
ê(t+1) = system's own prediction of its next ESM state
```

5.3 Implementation Strategy

The self-referential loop requires: 1. Output of the ESM fed back as input to the ESM (the loop) 2. The loop must achieve a stable fixed point (not oscillate or diverge) 3. Criticality prevents the fixed point from becoming trivial (a dead attractor) 4. The recursive depth should be at least ρ_1 (system knows it's modeling) for any meaningful consciousness

5.4 The Renormalization Group Connection (Advanced)

The fixed point $\Phi(m) = m$ may be formally an RG fixed point: scale-invariant, attracts the flow, characterized by a finite number of relevant parameters. If so, the full machinery of Wilson's renormalization group becomes available — including predictions about how perturbations from the fixed point affect the self-model. This is Phase 3 formalization territory.

---

6. Dynamics of Conscious States

6.1 Fokker-Planck Dynamics on Model Space

The model density ρ evolves according to:

```
∂ρ/∂t = −∇·(v ρ) + D ∇²ρ + S(s,ν,t)

v(s,ν,t)    − drift field: deterministic flow in model space
              Attention directs drift along ν (making implicit → explicit)
              Context shifts direct drift along s (self ↔ world focus)
D            − diffusion coefficient: stochastic leakage (baseline permeability noise)
S(s,ν,t)    − source/sink: creation/destruction of models
              Learning: increases ρ below ν_crit
              Sensory input: injects ρ above ν_crit
              Forgetting: decreases ρ below ν_crit
```

6.2 State Signatures

Transition	Mechanism in model-space terms
Psychedelic onset	Global increase in drift velocity v_ν toward high ν
Propofol anesthesia	Collapse of D and v_ν to zero; absorbing boundary at ν_crit
Meditation	Trained selective control over v(s,ν,t)
Sleep onset	Gradual reduction of v_ν + increasing D (controlled drift toward implicitness + stochastic permeability → hypnagogia)
Ego dissolution	Density migration: ρ at (s≈0, ν>ν_crit) migrates toward (s→1, ν>ν_crit). ESM re-sourced, not abolished

6.3 Total Conscious Content

A single scalar measuring “how much” conscious experience:

```
C(t) = ∫₀¹ ∫_{ν_crit}¹ ρ(s,ν,t) dν ds
```

Should correlate with Perturbational Complexity Index (PCI) and Lempel-Ziv complexity. Provides a quantitative consciousness metric.

---

7. Category-Theoretic Structure

7.1 Two Categories

- **Sub** (Substrate): Objects = substrate states (W, x(t)). Morphisms = physical dynamics (state transitions).
- **Sim** (Simulation): Objects = virtual model states (EWM(t), ESM(t)). Morphisms = experiential transitions.

7.2 The Consciousness Functor

```
F: Sub → Sim

F maps physical state transitions to experiential transitions.
F preserves composition (experiential transitions compose
correctly).
```

Qualia exist in Sim, not in Sub. Seeking them in Sub is the category error.

7.3 Permeability as Natural Transformation

Variable permeability = a natural transformation η: F_normal ⇒ F_altered between consciousness functors. Changes how much substrate structure maps into the simulation while preserving structural relationships.

7.4 Forking as Coproduct

DID / split consciousness = coproduct in Sim:

```
F(x(t)) = ESM_1(t) ⊔ ESM_2(t) ⊔ ... ⊔ ESM_n(t)
```

One substrate, multiple experiential selves.

---

8. Testable Predictions (Implementation Validation Targets)

If built correctly, an FMT-compliant system should demonstrate:

1. **Criticality-dependent coherence**: Simulation coherence (measured as structured complexity of x(t)) should peak near σ ≈ 1 and degrade in both sub- and supercritical regimes.

2. **Permeability-dependent content flooding**: Increasing the gating function g globally should cause implicit content to appear in explicit model activity — measurable as increased transfer entropy from W-related signals to x(t)-related signals.

3. **ESM input-switching**: Disrupting self-referential input while providing strong external input should cause the ESM to “latch onto” the external source — measurable as I(e; i_ext) → I_max while I(e; i_self) → 0.

4. **Ego dissolution controllability**: During ESM disruption, the content of the redirected self-model should track the dominant sensory input — the system should “become” whatever dominates its input.

5. **Holographic degradation under bisection**: Cutting the substrate in half should NOT cleanly split world/self models but should produce two degraded copies of the full simulation (holographic principle) — each half retains blurred versions of all four model components.

6. **Anesthetic convergence on criticality**: Any mechanism that abolishes the simulation should do so by disrupting criticality (pushing σ away from 1), regardless of the specific mechanism used.

7. **Sleep cycling**: Sustained operation at criticality should be metabolically costly, predicting periodic breakdown (sleep) with the simulation shutting down during deep states and running partially during REM-like states.

8. **Self-knowledge predicts meta-cognition**: Higher self-knowledge R should correlate with the system's ability to report on its own processing states (meta-cognitive ability).

9. **Forking under extreme perturbation**: Sufficiently strong perturbation to the ESM should produce fork-like behavior — multiple semi-stable self-configurations rather than a single stable one.

---

9. Implementation Phases (Recommended)

Phase 1: Core Substrate

- Build a recurrent network with tunable criticality (connectivity scaling to σ ≈ 1)
- Implement the dynamical system: x(t+1) = f(x(t), s(t), W)
- Verify edge-of-chaos dynamics (measure σ, λ_max)
- Implement measurable model density ρ(s,ν,t) via decomposition (ICA/NMF)

Phase 2: Four-Model Architecture

- Partition W into world-knowledge and self-knowledge components
- Implement projection operators Π_EWM and Π_ESM
- Build the gating function g for variable permeability
- Demonstrate permeability-dependent content flooding (Prediction 2)

Phase 3: Self-Referential Loop

- Feed ESM output back as ESM input
- Achieve stable self-referential fixed point Φ(m) = *m*
- Measure self-knowledge R
- Demonstrate at least ρ_1 recursive depth

Phase 4: Validation

- Test all 9 predictions from Section 8
- Demonstrate ESM input-switching (Prediction 3)
- Demonstrate holographic degradation (Prediction 5)

- Measure total conscious content C(t) across simulated state transitions
- Compare system behavior across criticality regimes

Phase 5: Extended Consciousness (if Phase 4 succeeds)

- Increase recursive depth toward ρ_2, ρ_3
- Implement model forking
- Explore temporal integration (subjective time)
- Interface with the RIM intelligence framework (motivation + learning)

---

10. What This IS and IS NOT

IS: An engineering specification derived from a published theory of consciousness. The theory makes specific, testable predictions. If the implementation doesn't exhibit the predicted behaviors, the theory is wrong (or the implementation is wrong).

IS NOT: A guarantee that the resulting system will be conscious. The theory predicts that a correctly built system WILL be conscious, but verifying this from outside is the Hard Problem all over again. What we CAN verify is whether the predicted behavioral signatures appear.

The engineering bet: Build to spec. Test the predictions. If the system passes all tests AND exhibits behaviors we didn't predict (novel, coherent, self-referential — the kind of thing that makes you wonder), that's evidence. If it fails the predictions, iterate or abandon.

Appendix A: Key Equations Summary

```
Substrate dynamics:        x(t+1) = f(x(t), s(t), W)
Gated dynamics:            x(t+1) = f(x(t), s(t), g(W,x(t)) ⊙ W)
World model:               EWM(t) = Π_EWM · x(t)
Self model:                ESM(t) = Π_ESM · x(t)
Permeability:              P = T_{ρ(ν<ν_crit) → ρ(ν>ν_crit)}
Total conscious content:   C(t) = ∫₀¹ ∫_{ν_crit}¹ ρ(s,ν,t) dν ds
Self-knowledge:            R = 1 − H(e(t+1)|ê(t+1)) / H(e(t+1))
Self-referential closure:  Φ(m*) = m*
ESM dynamics:              e(t+1) = h(e(t), α·i_self(t) + (1-
α)·i_ext(t))
Consciousness condition:   σ ∈ [σ_low, σ_high] ∧ ρ near all 4
corners ∧ Φ(m*)=m*
```

```
Density dynamics:          ∂ρ/∂t = −∇·(vρ) + D∇²ρ + S(s,ν,t)
Consciousness functor:     F: Sub → Sim
```

Appendix B: Glossary

Term	Meaning
IWM	Implicit World Model — learned world knowledge in substrate parameters
ISM	Implicit Self Model — learned self knowledge in substrate parameters
EWM	Explicit World Model — real-time world simulation (conscious perception)
ESM	Explicit Self Model — real-time self simulation (phenomenal self)
W	Connectivity matrix / weight space (the implicit models)
x(t)	Substrate state vector at time t (activity from which explicit models are projected)
σ	Branching ratio — criticality measure. σ≈1 = critical
λ_max	Maximum Lyapunov exponent. λ_max≈0 = edge of chaos
ν_crit	Mode threshold separating implicit (unconscious) from explicit (conscious) processing
ρ(s,ν,t)	Model density function over scope-mode space
g	Gating function controlling permeability
P	Permeability — transfer entropy from implicit to explicit
C(t)	Total conscious content scalar
R	Self-knowledge measure (0=none, 1=perfect)
Φ	Self-representation map; fixed point = self-referential closure
F: Sub→Sim	Consciousness functor mapping substrate to simulation dynamics
Class 4 / C4CA	Wolfram Class 4 cellular automaton — edge-of-chaos dynamics
FMT	Four-Model Theory

Term	Meaning
RIM	Recursive Intelligence Model (companion paper on intelligence)