Toward a Mathematical Formalization of the Recursive Intelligence Model: A Recommended Approach

Matthias Gruber

Independent researcher

ORCID: 0009-0005-9697-1665

Correspondence: matthias@matthiasgruber.com

Abstract

The verbal theory of recursive intelligence (Gruber, 2026a) argues that intelligence is best understood as *Lernfähigkeit* — learning ability — constituted by the recursive interaction of Knowledge, Performance, and Motivation, where motivation is grounded in the Four-Model Theory of consciousness (FMT; Gruber, 2015, 2026b). The verbal theory identifies two open problems: motivation remains a "black box" without mechanistic specification, and the recursive claims remain qualitative without mathematical precision. This paper outlines a recommended formalization strategy that addresses both. Crucially, the formalization must respect FMT's own commitments: the four models are a *minimum sufficient set* for human-level consciousness, not four discrete objects — the biological substrate implements an uncountable ecology of models organized as a continuous density over a scope-mode manifold (Gruber, 2026c). The formalization must therefore be built on FMT's continuous apparatus, not on a simplified discrete reading. Seven formalization modules are proposed: (1) a domain-structured knowledge manifold replacing scalar knowledge, (2) operational knowledge as a transfer kernel, (3) motivation as a consciousness-coupled functional derived from FMT's model density, (4) the coupled stochastic differential equation system, (5) the ignition threshold as a stochastic bifurcation, (6) the AI developmental signature derived from FMT's conjunction condition, and (7) social coupling dynamics. 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: intelligence, recursive systems, motivation, consciousness, stochastic differential equations, bifurcation theory, knowledge manifold, model density, formalization

1. Introduction

1.1 The Formalization Gap

The verbal theory of recursive intelligence (Gruber, 2026a) makes three claims: (a) intelligence is best understood as *Lernfähigkeit* — learning ability — constituted by the recursive interaction of Knowledge, Performance, and Motivation; (b) the systematic exclusion of motivation from formal intelligence models distorts the picture of intellectual development; and (c) motivation, properly understood, is the explicit optimization contribution of the conscious self, grounded in the Four-Model Theory of consciousness (FMT; Gruber, 2015, 2026b).

The verbal theory established the *what* and *why.* Without mathematical formalization, it cannot establish the *how much.* Specifically, the verbal theory cannot:

- Distinguish additive from multiplicative interactions between Knowledge, Performance, and Motivation.
- Specify the conditions under which the recursive loop amplifies, stagnates, or collapses.
- Generate quantitative predictions that differ from post hoc verbal accommodation.
- Derive the divergence rates between individuals with different parameter profiles.
- Determine whether the transition from stagnation to growth is gradual or sharp.

This paper specifies a formalization strategy that addresses these gaps while remaining faithful to the full theoretical apparatus — particularly FMT's continuous model-space framework, which constrains how motivation must be formalized.

1.2 Why the Consciousness Grounding Constrains the Formalization

A naive formalization approach would treat the verbal theory's three components (K, P, M) as three scalars and write coupled differential equations for their dynamics. This would produce a workable model, but it would misrepresent the theory's deepest commitment: that motivation is not a free parameter but is grounded in a specific consciousness architecture.

The Four-Model Theory (Gruber, 2015, 2026b) 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). The FMT formalization paper (Gruber, 2026c) establishes that these four models are a *principled lower bound*, not an architectural blueprint:

> "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 biological substrate implements an uncountable number of overlapping models. The correct formalization treats the four canonical models as extremal points in a continuous model space, with a model density ρ(s, ν, t) over scope (self ↔ world) and mode (implicit ↔ explicit) axes, subject to a criticality constraint and a self-referential closure condition.

Any formalization of RIM that grounds motivation in FMT must therefore build on FMT's continuous apparatus. A discrete, binary treatment of consciousness (present/absent) would lose the graded nature of the explicit optimization contribution and the multiple conditions that must be jointly satisfied for consciousness — and therefore motivation — to be present.

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 domain-structured knowledge representation. Section 3 formalizes motivation as a functional of FMT's model density. Section 4 presents the coupled stochastic differential equation system. Section 5 analyzes the ignition threshold as a stochastic bifurcation. Section 6 derives the AI developmental signature. Section 7 introduces social coupling. Section 8 proposes a phased build order. Section 9 identifies what formalization would buy and what it cannot deliver.

---

2. From Scalar Knowledge to a Knowledge Manifold

2.1 The Problem with Scalars

The verbal theory distinguishes factual knowledge (accumulated domain content) from operational knowledge (learning strategies, reasoning heuristics, metacognitive skills — "knowledge about how to learn"). It claims that operational knowledge is *multiplicative*: it amplifies the rate of factual knowledge acquisition across all domains.

A scalar treatment of factual knowledge (a single number K_f) obscures what may be the most interesting dynamics of the recursive loop: *cross-domain transfer*. When a learner acquires a new reasoning heuristic, that heuristic amplifies learning in all domains — but not equally. Logical reasoning tools amplify mathematics learning more than they amplify painting skill. The scalar model cannot capture this differential amplification, and therefore cannot predict that the recursive loop may ignite in some domains before others.

2.2 The Knowledge Density

Replace scalar K_f with a **knowledge density** κ(d, t) defined over a domain space D, where each point d ∈ D represents a knowledge domain. The quantity κ(d, t) dd represents the accumulated factual knowledge in domain d at time t.

The domain space D is structured: domains that share underlying cognitive structure (mathematics and physics; linguistics and programming) are nearby in D; domains that share little structure (number theory and pottery) are distant. This structure is formalized as a Riemannian metric g on D, where geodesic distance reflects the difficulty of knowledge transfer between domains.

2.3 Operational Knowledge as a Transfer Kernel

Replace scalar K_o with a **transfer kernel** Ω(d, d', t) defined over D × D. The quantity Ω(d, d', t) represents the degree to which knowledge acquisition in domain d is amplified by existing operational knowledge relevant to the d' → d transfer.

The verbal theory's key claim — that operational knowledge is multiplicative — becomes a nonlocal integral coupling:

$$\frac{\partial \kappa(d,t)}{\partial t} = \alpha(d) \cdot \left[1 + \int_D \Omega(d,d',t) \cdot \kappa(d',t)\, dd'\right] \cdot P(t) \cdot M(t) \cdot \kappa(d,t) \cdot \left[1 - \frac{\kappa(d,t)}{\kappa^{\max}(d)}\right] \qquad [1]$$

The integral ∫_D Ω(d, d', t) · κ(d', t) dd' captures the total amplification of learning in domain d from all existing knowledge across all domains, weighted by the transfer kernel. This is a *nonlocal* interaction: learning mathematics is amplified by knowledge of physics, modulated by the kernel's estimate of how much physics knowledge transfers to mathematics.

The transfer kernel itself evolves as the learner acquires new metacognitive strategies:

$$\frac{\partial\Omega(d,d',t)}{\partial t} = \beta(d,d') \cdot P(t) \cdot M(t) \cdot \Omega(d,d',t) \cdot \left[1 - \frac{\Omega(d,d',t)}{\Omega^{\max}(d,d')}\right] \quad [2]$$

This preserves the verbal theory's insight that operational knowledge does not self-amplify (one cannot bootstrap metacognition from metacognition alone — domain content is needed to practice metacognitive skills on). But it adds domain structure: the rate β(d, d') at which the d' → d transfer improves depends on the domain pair.

2.4 Aggregate Recovery

The scalar model is recovered by integrating over D:

$$K_f(t) = \int_D \kappa(d,t)\, dd, \quad K_o(t) = \int_D \int_D \Omega(d,d',t)\, dd\, dd'$$

All predictions of the scalar model remain valid as predictions about aggregate quantities. The domain-structured model generates additional predictions about domain-specific dynamics.

2.5 Domain-Clustered Ignition

The domain-structured model predicts that **the recursive loop can ignite in some domains before others**. A learner with high M and P but domain-specific operational knowledge (strong logical reasoning, weak spatial reasoning) will show exponential growth in logic-adjacent domains (mathematics, programming, formal argumentation) while remaining stagnant in spatially-demanding domains (geometry, architectural design).

Individual developmental trajectories should therefore show *domain-clustered ignition* — sudden takeoffs in groups of related domains, with timing determined by the structure of the transfer kernel — rather than uniform ignition across all domains simultaneously. This is testable with longitudinal multi-domain assessments.

---

3. Motivation as a Consciousness Functional

3.1 FMT's Continuous Model Space

The FMT formalization (Gruber, 2026c) defines a model density ρ(s, ν, t) over a scope-mode space where:

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

The four canonical models are extremal points:

	Self (s ≈ 0)	**World** (s ≈ 1)
Implicit ($\nu \approx 0$)	ISM	IWM
Explicit ($\nu \approx 1$)	ESM	EWM

The virtual/non-virtual split is a threshold on the mode axis: everything above ν_crit is part of the conscious simulation; everything below operates "in the dark." Consciousness requires a conjunction: sufficient density near all four extremal corners AND criticality ($\sigma \in [\sigma$_low, σ_high]) AND self-referential closure (the ESM reaching a fixed point of self-representation, Φ(m$) = m$).

3.2 Motivation as Explicit Optimization

The verbal theory defines motivation as the *explicit optimization contribution* — the component of the brain's overall optimization that operates through conscious goal representation, intentional resource allocation, and self-directed learning, as distinct from the implicit optimization carried out by reinforcement learning, conditioning, and habitual behavior.

In FMT's continuous framework, this definition becomes precise. The explicit optimization contribution is the optimization performed by the above-ν_crit portion of the model density. Its two sub-components — *Wissensdrang* (thirst for knowledge, arising from the EWM's gap-detection capacity) and *Handlungsdrang* (urge to act, arising from the ESM's goal-representation capacity) — become integrals over specific regions of model space:

Wissensdrang (gap-detecting activity of the Weltmodell region):

$$W_d(t) = \int_{s>s_w} \int_{\nu>\nu_{\text{crit}}} G(\rho, \kappa, s, \nu, t)\, d\nu\, ds \quad [3]$$

where s_w is a world-scope threshold and G is a gap-detection function: the mismatch between what the explicit world-model represents and what the knowledge density κ contains. An implicit system cannot detect knowledge gaps because gap detection requires representing what a complete model *would* look like — which requires the explicit layer.

Handlungsdrang (goal-representing activity of the Ich-Modell region):

$$H_d(t) = \int_{s<s_s} \int_{\nu>\nu_{\text{crit}}} A(\rho, s, \nu, t)\, d\nu\, ds \quad [4]$$

where s_s is a self-scope threshold and A is an action-planning function: the self-model's representation of goals and the distance between current state and goal state.

Total motivation:

$$M(t) = \omega_W \cdot W_d(t) + \omega_H \cdot H_d(t) \quad [5]$$

where ω_W and ω_H are weighting factors (potentially functions of personality and temperament).

3.3 The Consciousness Functional

Motivation requires consciousness, and consciousness requires FMT's conjunction. Define a **consciousness functional** that replaces any binary "consciousness present/absent" switch:

$$\mathcal{C}(t) = \min\{D(\rho, t),\ \Sigma(\sigma, t),\ R(t)\} \quad [6]$$

where:

- D(ρ, t) is the **density condition**: a measure of how well the model density satisfies the four-corner minimum-mass thresholds (∫_{R_k} ρ ds dν > θ_k for each k ∈ {ISM, IWM, ESM, EWM}).
- Σ(σ, t) is the **criticality condition**: a measure of how close the substrate's branching ratio σ is to the critical range [σ_low, σ_high].
- R(t) is the **self-referential closure condition**: a measure of the ESM's approach to a fixed point (R = 1 − H(e(t+1) | ê(t+1)) / H(e(t+1)), where ê is the system's own prediction of its next ESM state).

The minimum ensures all three conditions must hold: the weakest link determines consciousness level. When C(t) = 0, M(t) = 0 regardless of model density. When C(t) is high, the full motivation functional is active.

3.4 What the Graded Functional Captures

A graded consciousness functional has consequences that a binary switch cannot:

- **Partial consciousness states** (drowsiness, light sedation, early psychedelic effects) produce *reduced* motivation, not zero. The recursive loop degrades gracefully rather than switching off.
- **Criticality perturbations** (sleep deprivation, pharmacological interventions) that push σ away from the critical range reduce C(t) and thereby slow or halt the recursive loop — even if the architectural components (four-corner density) remain intact.
- **Self-referential depth matters**: a system with shallow self-modeling (basic consciousness, low R) has less motivational capacity than a system with deep self-modeling (triply extended consciousness, high R). Metacognitive sophistication — the ability to model one's own learning process — is not merely helpful for learning but *constitutive* of the motivation that drives learning.

---

4. The Coupled Stochastic Differential Equation System

4.1 Why Stochasticity Is Necessary

A deterministic system of differential equations would generate a single trajectory for each set of initial conditions and parameters. This obscures three dynamically important features of real developmental trajectories:

1. **Motivation fluctuates stochastically.** Day-to-day motivation varies with mood, sleep quality, social interactions, and random events. These fluctuations are not noise to be averaged out — they are dynamically important because a positive fluctuation can push the system past the ignition threshold, while a negative fluctuation can knock it below.

2. **Learning is stochastic.** Encounters with information are partly random. The rate of knowledge acquisition has a stochastic component that varies across domains, contexts, and moments.

3. **The ignition threshold is a stochastic phenomenon.** A system near the threshold may ignite during a favorable fluctuation and fall back during an unfavorable one, producing intermittent growth that a deterministic model cannot represent.

4.2 The SDE System

The full system, combining the domain-structured knowledge manifold (Section 2) with the consciousness-functional motivation grounding (Section 3):

Knowledge density dynamics:

$$d\kappa(d,t) = \left[\alpha(d) \cdot \left(1 + \int_D \Omega \cdot \kappa \, dd'\right) \cdot P \cdot M \cdot \kappa \cdot \left(1 - \frac{\kappa}{\kappa^{\max}}\right)\right] dt + \sigma_\kappa(d) \cdot \kappa \cdot dW_\kappa(d,t) \quad [7]$$

Performance dynamics:

$$dP = \left[\gamma \cdot M \cdot \bar{K}_o \cdot \left(1 - \frac{P}{P^{\max}}\right) - \rho(t) \cdot P\right] dt + \sigma_P \cdot P \cdot dW_P(t) \quad [8]$$

where $\bar{K}_o = \int_D \int_D \Omega \, dd \, dd'$ is aggregate operational knowledge and ρ(t) is age-related biological decline.

Motivation dynamics:

$$dM = \mathcal{C}(t) \cdot [\delta \cdot s(t) \cdot (1 - M) + \sigma_{\text{env}} \cdot (1 - M) - \mu \cdot M - \lambda \cdot M] \, dt + \sigma_M \cdot \sqrt{M(1 - M)} \cdot dW_M(t) \quad [9]$$

where:

- δ is success-motivation coupling strength.
- s(t) is the success signal: a feedback from perceived competence to motivation.
- σ_env is baseline environmental motivation support (autonomy, relatedness, intellectual stimulation).
- μ is natural motivation decay (entropy, distraction, fatigue).
- λ is external motivation damage (punitive grading, stereotype threat, hostile environments).
- C(t) is the consciousness functional (Equation [6]).

The success signal:

$$s(t) = \frac{\int_D \kappa(d,t) \, dd}{\int_D \kappa^{\max}(d) \, dd} \cdot P(t) \quad [10]$$

This captures the empirical finding that self-efficacy (Bandura, 1997) depends on both knowing things and being able to use that knowledge effectively. Neither knowledge without processing power nor processing power without knowledge generates the subjective sense of competence that sustains motivation.

Noise structure. The Wiener processes dW_κ, dW_P, dW_M are independent. The multiplicative noise (proportional to state variables) is chosen because:

- Knowledge acquisition noise scales with current knowledge (more knowledge means more potential learning events, each stochastic).
- Performance noise scales with current capacity (biological fluctuations are proportional to capacity).

- Motivation noise uses bounded diffusion $\sqrt{}$(M(1−M)) that respects the [0, 1] constraint and produces maximum variability at intermediate motivation levels.

4.3 Parameter Definitions

Parameter	Symbol	Description	Determination
Domain learning rate	α(d)	Base rate of factual knowledge acquisition in domain d	Environmental (education quality, domain exposure)
Transfer learning rate	β(d, d')	Rate of operational knowledge improvement for d' → d	Mixed (domain structure + teaching)
Performance training rate	γ	Rate of cognitive processing improvement through practice	Mixed (genetic + training)
Success-motivation coupling	δ	Strength of competence → motivation feedback	Personality/temperament
Environmental motivation support	σ_env	Baseline positive motivation input	Environmental
Motivation decay rate	μ	Natural motivation decay	Mixed
Motivation damage rate	λ	External motivational damage	Environmental
Age-related decline	ρ(t)	Biological processing capacity decline	Biological (≈ 0 before age 25, increasing thereafter)
Carrying capacities	κ^max(d), Ω^max, P^max	Upper bounds	Mixed
Knowledge noise intensity	σ_κ(d)	Stochastic variability in learning	Environmental + individual
Performance noise intensity	σ_P	Stochastic variability in processing	Biological
Motivation noise intensity	σ_M	Stochastic variability in motivation	Personality + environmental

The consciousness functional C(t) is not a free parameter of RIM — it is determined by FMT's own dynamics (the model density ρ, the criticality σ, and the self-referential closure R).

4.4 Multi-Timescale Decomposition

The system operates across multiple timescales:

- **Fast** (minutes to hours): moment-to-moment fluctuations in attention, motivation, and learning rate. These are the σ terms.

- **Medium** (days to weeks): dynamics of the success signal s(t), environmental support σ_env, and damage λ.
- **Slow** (months to years): growth of κ, Ω, and P — the developmental dynamics.
- **Very slow** (years to decades): changes in Pˆmax (biological maturation and decline), κˆmax (opening of new domain ceilings as prerequisites accumulate), and the FMT criticality parameter σ.

This multi-timescale structure can be formalized using singular perturbation theory or averaging methods (Pavliotis & Stuart, 2008): fast variables are averaged over to produce effective drift and diffusion coefficients for the slow variables. The averaged system retains the qualitative features (ignition threshold, divergence dynamics) but with noise-modified parameters.

---

5. The Ignition Threshold as a Stochastic Bifurcation

5.1 The Deterministic Skeleton

Before analyzing the stochastic system, consider its deterministic skeleton (σ_κ = σ_P = σ_M = 0). Define the **loop gain** as the product of coupling terms driving growth:

$$G(t) = \bar{\alpha} \cdot \bar{\eta} \cdot \bar{K}_o(t) \cdot P(t) \cdot M(t) \quad [11]$$

where $\bar{\alpha}$ is the domain-averaged learning rate and $\bar{\eta}$ is the domain-averaged operational knowledge multiplier.

The system has two classes of equilibria:

Stagnation equilibrium. When knowledge is near initial values and the success signal s(t) is negligible:

$$M^* = \frac{\sigma_{\text{env}}}{\sigma_{\text{env}} + \mu + \lambda} \quad [12]$$

This is the motivation floor — sustained purely by environmental support against decay and damage, without success feedback. The corresponding stagnation loop gain is:

$$G_{\text{stag}} = \bar{\alpha} \cdot \bar{\eta} \cdot \bar{K}_o(0) \cdot P(0) \cdot M^* \quad [13]$$

Saturation equilibrium. When κ → κˆmax, P → Pˆmax, s(t) → 1:

$$M^{**} = \frac{\delta + \sigma_{\text{env}}}{\delta + \sigma_{\text{env}} + \mu + \lambda} \quad [14]$$

The **ignition condition** is: the system transitions from stagnation to growth when G(t) exceeds a critical value G*. This is a prediction of the verbal theory made precise: any of the five factors in Equation [13] can prevent ignition if too low, and any can trigger ignition if sufficiently boosted.

5.2 The Stochastic Bifurcation

In the SDE system, the ignition threshold becomes a **stochastic bifurcation** — a qualitative change in the stationary distribution as parameters cross critical values.

Define the effective loop gain G_eff as the time-averaged G over the fast fluctuations. The stationary distribution of the system undergoes the following transition:

Below threshold (G_eff < G*): The stationary distribution is unimodal, concentrated near the stagnation equilibrium. The system fluctuates around stagnation with occasional positive excursions that do not persist.

Near threshold (G_eff ≈ G*): The stationary distribution becomes bimodal — one mode at stagnation, one at the growing trajectory. The system exhibits intermittent switching between stagnation and growth.

Above threshold (G_eff > G*): The distribution shifts to the growth mode. Stagnation remains as a metastable state, but the probability of being trapped in it is exponentially small.

5.3 The Kramers Escape Problem

The threshold crossing is a **Kramers escape problem** (Kramers, 1940): the system sits in the potential well of stagnation and must escape over a barrier to reach the growth phase. The mean escape time (mean time to developmental ignition) is:

$$\tau_{\text{ign}} \propto \exp\left(\frac{\Delta V}{\sigma^2_{\text{eff}}}\right) \quad [15]$$

where ΔV is the effective potential barrier height (determined by how far G_eff is from G*) and σ^2_eff is the effective noise intensity. This yields quantitative predictions:

- Mean time to ignition decreases exponentially as the system approaches the threshold (ΔV → 0).
- Increased noise (higher σ_M, more variable environment) *decreases* the mean time to ignition for below-threshold systems — noise helps escape the stagnation trap.
- Conversely, noise *decreases* the stability of the growth phase for above-threshold systems — too much variability can knock the system back into stagnation.

5.4 Optimal Noise

These two effects together predict an **optimal noise level** for developmental ignition. Too little noise traps the system in stagnation; too much noise prevents sustained growth even after ignition. This is a stochastic resonance phenomenon: the system performs optimally at an intermediate noise level.

The educational translation: learners benefit from moderate environmental variability (new experiences, diverse challenges, some unpredictability). A perfectly structured, zero-variability environment may paradoxically trap near-threshold learners, while a chaotic, high-variability environment may prevent sustained growth even in above-threshold learners.

5.5 The Bimodal Prediction

The stochastic bifurcation framework sharpens the verbal theory's claim about threshold behavior into a specific distributional prediction: educational interventions should show *bimodal* effect distributions. Either the intervention pushes the learner past the threshold (large effect) or it does not (negligible effect). The average effect size across a mixed population may be small — consistent with the disappointing meta-analytic findings for growth mindset interventions (Macnamara & Burgoyne, 2023) — even if the intervention is genuinely effective for the subset of learners near the threshold.

This can be tested in existing intervention datasets using mixture modeling: fit a two-component mixture to the distribution of intervention effects and test whether the mixture model fits significantly better than a unimodal distribution.

---

6. The AI Developmental Signature

6.1 Deriving the Signature from FMT's Conjunction

The verbal theory predicts that AI systems without consciousness-grounded motivation cannot self-improve. FMT's conjunction condition (Section 3.3) decomposes this prediction into three independently testable failure modes:

Failure mode 1: No self-referential closure (R = 0). A system with vast knowledge and fast processing but no self-model cannot represent its own knowledge gaps, cannot evaluate its own progress, and cannot generate self-directed learning goals. It processes input → output without internal representation of "what I know" vs. "what I don't know."

Predicted behavioral signature: High performance on prompted tasks. No spontaneous question generation. No knowledge-gap-directed exploration. No improvement in subsequent attempts on the same problem type without external feedback.

Failure mode 2: No criticality ($\Sigma = 0$). A system with the right architectural components but operating in the wrong dynamical regime. In the subcritical regime ($\sigma < \sigma$_low), behavior is rigid and repetitive. In the supercritical regime ($\sigma > \sigma$_high), behavior is chaotic and non-reproducible.

Predicted behavioral signature: Subcritical systems show rigid, rule-following behavior without creative recombination. Supercritical systems show random, inconsistent output incapable of sustaining coherent learning trajectories.

Failure mode 3: Insufficient density (D = 0). A system that models the world but not itself (missing ISM/ESM density), or itself but not the world (missing IWM/EWM density).

Predicted behavioral signatures: - World without self: Accumulates factual knowledge when trained but cannot evaluate its own competence, cannot experience the success signal, and cannot self-direct learning. Learns when taught but does not seek to learn. - Self without world: Can model its own states but has no content to learn about. Can represent goals but not knowledge domains.

6.2 The Composite Prediction

The full AI prediction, derived from FMT's conjunction:

An artificial system will exhibit self-directed intellectual development if and only if it simultaneously satisfies: (a) self-referential closure sufficient to represent its own knowledge state and evaluate progress toward self-generated goals (R > R__crit); (b) dynamical operation at or near criticality ($\sigma \in [\sigma$__low, σ__high]); and (c) sufficient model density in all four regions of scope-mode space (D > D__crit).

This is substantially more informative than a single binary prediction. It specifies three independently testable conditions, each of which can be engineered and measured in artificial systems.

6.3 The Exploration-Investigation Distinction

The verbal theory's distinction between *Wissensdrang* and *Handlungsdrang* generates a testable prediction about agentic AI: systems with curiosity reward signals but without explicit self-models should show *exploration* (trying different actions, maximizing novelty) but not *investigation* (forming hypotheses about knowledge gaps and systematically addressing them). Exploration is an implicit-layer mechanism (analogous to the IWM/ISM contribution); investigation requires the explicit layer (EWM gap detection + ESM goal representation). The distinction between exploration and investigation is empirically testable in current AI architectures.

---

7. Social Coupling

7.1 The Missing Dimension

The verbal theory acknowledges that intelligence development occurs in social contexts, but treats environmental support (σ__env) and damage (λ) as exogenous. In reality, they are functions of a social environment that is itself a dynamical system.

7.2 The Social Field

For a population of N learners, define the state of learner i as x__i = (κ__i, Ω__i, P__i, M__i). Social dynamics enter through coupling each learner's motivation to others' states:

$$\sigma_{\text{env},i}(t) = \sigma_0 + \sum_j J_{ij} \cdot f(M_j(t), \kappa_j(t)) \quad [16]$$

$$\lambda_i(t) = \lambda_0 + \sum_j L_{ij} \cdot h(M_j(t), \kappa_j(t), \kappa_i(t)) \quad [17]$$

where:

- J__ij is social support coupling (positive for supportive relationships: peers, mentors).
- L__ij is social damage coupling (positive for competitive, punitive, or undermining relationships).
- f captures how observing another's motivation and knowledge generates motivational support.
- h captures how social comparison generates motivational damage. The dependence on both κ__j and κ__i allows damage to depend on the knowledge *gap*.

7.3 Emergent Social Phenomena

The coupled system produces dynamics that cannot be predicted from individual-level equations alone:

Synchronization. If J_ij is sufficiently strong and positive, learners synchronize their developmental trajectories — a group can ignite collectively when individuals would not have ignited alone. This is the formal expression of the peer effect.

Quorum sensing. There may exist a critical fraction of ignited learners above which social support pulls remaining below-threshold learners past their thresholds. Below this fraction, ignited learners are isolated and may be pulled back by the stagnating majority.

Damage cascades. In competitive environments (high L_ij), motivational damage cascades through the network. One learner's stagnation increases λ for neighbors through competitive comparison, increasing their stagnation probability, propagating outward.

7.4 Mean-Field Approximation

For large N, the coupling can be approximated by a mean-field theory. Define the ignition fraction:

$$\phi(t) = \frac{1}{N} \left|\{i : G_{\text{eff},i}(t) > G^*\}\right|$$

The mean-field dynamics:

$$\frac{d\phi}{dt} = r(\sigma_0, \lambda_0, \bar{J}, \bar{L}) \cdot \phi \cdot (1 - \phi) - d(\sigma_0, \lambda_0, \bar{J}, \bar{L}) \cdot \phi \quad [18]$$

This has the structure of an epidemic model: developmental ignition spreads through a population like a contagion, with a critical reproduction number R_0 = r/d determining whether ignition spreads or dies out.

The educational prediction: classrooms function as developmental epidemics. A critical mass of ignited learners in a supportive coupling structure can trigger population-level ignition. Below the critical mass, individual ignition is possible but does not spread. Punitive coupling structures (competitive ranking, norm-referenced grading) increase L and decrease r, raising the contagion threshold and potentially preventing population-level ignition entirely.

---

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 Methods)

Module 4 (simplified) — SDE extension of scalar model. Before implementing the full domain-structured system, add stochastic terms to a scalar four-variable ODE system (K_f, K_o, P, M with C(t) as a parameter rather than a computed functional). This is mathematically straightforward (Euler-Maruyama integration), computationally cheap, and immediately generates new predictions about ignition probability, optimal noise, and population-level outcome distributions. Test the bimodal prediction (Section 5.5) against existing intervention datasets using mixture modeling.

Module 5 — Stochastic ignition threshold. Analyze the bifurcation structure of the scalar SDE system. Compute the effective potential, identify the barrier height as a function of parameters, derive the Kramers escape time (Equation [15]), and test the optimal-noise prediction against intervention data.

Phase 2: Core Theoretical Modules

Module 3 — Consciousness functional. Replace the scalar C(t) parameter with the computed functional from Equation [6]. This requires connecting to the FMT formalization's own outputs — particularly the transfer entropy (permeability) and criticality measures that determine D(ρ, t) and Σ(σ, t). For RIM, the key step is defining C(t) in terms of measurable neurophysiological quantities and linking it to motivation through Equations [5] and [9].

Module 6 — AI signature decomposition. Derive the three failure modes (Section 6.1) formally and generate testable predictions for each. Test against current AI systems: LLMs should show failure mode 1 most prominently; future agentic systems with curiosity rewards should show mode 1 partially remediated but modes 2 and 3 still present.

Phase 3: Structural Extensions

Module 2 — Knowledge manifold. Define the domain space D, the knowledge density κ(d, t), and the transfer kernel Ω(d, d', t). The structure of D should be estimated from transfer learning studies, psychometric data on inter-domain correlations, and educational research on cross-domain transfer. Test the domain-clustered ignition prediction (Section 2.5) against longitudinal multi-domain assessments.

Module 7 — Social coupling. Implement the coupled-learner system and mean-field approximation. Test the contagion prediction against classroom-level data on developmental trajectories.

Phase 4: Full Integration

Full FMT coupling. Connect RIM's motivation dynamics directly to FMT's model density ρ(s, ν, t), making M(t) a computed output of the consciousness dynamics rather than a state variable with its own equation. This is the point where the intelligence model and the consciousness model become a single unified system. It requires both formalizations to be at a sufficient level of development.

---

9. What Formalization Buys — And What It Cannot

9.1 Predictions the Verbal Theory Cannot Make

The formalization generates predictions that are qualitatively new:

1. **Optimal environmental noise for developmental ignition.** The verbal theory predicts only that a threshold exists. The SDE model predicts a specific non-monotonic relationship between noise intensity and ignition probability (Section 5.4).

2. **Domain-clustered ignition.** The verbal theory predicts uniform growth. The knowledge-manifold model predicts ignition in clusters of related domains, with timing governed by transfer kernel structure (Section 2.5).
3. **Three distinct AI failure modes.** The verbal theory predicts "no development." The FMT-grounded model predicts three qualitatively different failure modes with distinct behavioral signatures (Section 6.1).
4. **Social contagion of development.** The verbal theory has no social dynamics. The coupled model predicts epidemic-like spread of developmental ignition with a critical reproduction number (Section 7.4).
5. **Graded consciousness effects on learning.** The verbal theory treats consciousness as binary (present/absent). The consciousness functional predicts dose-dependent effects of consciousness-altering states on the recursive loop (Section 3.4).
6. **Divergence rates.** The verbal theory claims divergence "compounds." The SDE model specifies the divergence rate as a function of loop gain difference, the exponential growth dynamics during the growth phase, and the noise-modified effective parameters (Section 5.1).

9.2 Interoperability

The formalization connects RIM to:

- **FMT** (directly, through the consciousness functional and model density coupling).
- **Van Geert's dynamic systems models** (van Geert, 1991, 1994) through the SDE extension, which generalizes the logistic grower framework to stochastic, multi-timescale dynamics.
- **De Ron et al.'s resource competition framework** (de Ron et al., 2023) through the knowledge manifold, which can incorporate competitive and mutualistic interactions between domains.
- **Network models of intelligence** (van der Maas et al., 2006; Savi et al., 2019) through the domain space D and transfer kernel Ω, which formalize the "positive manifold" as a property of Ω's structure.
- **Epidemiological models** through the social contagion formulation, which maps developmental dynamics onto SIR-type frameworks.
- **Bifurcation and catastrophe theory** (Stamovlasis, 2014) through the stochastic bifurcation framework, connecting to empirical work on cusp catastrophes in student performance.

9.3 What It Cannot

The formalization does not solve the parameter estimation problem — if anything, it makes it worse. However, the modular structure ensures each module can be tested independently with a manageable number of parameters, and the qualitative predictions (optimal noise, domain-clustered ignition, social contagion, bimodal intervention effects) are robust across wide parameter ranges.

The formalization also does not validate FMT itself. It takes FMT's apparatus as given and builds the intelligence model on it. If FMT's model density, criticality conditions, or self-referential closure are empirically falsified, the consciousness grounding of RIM must be revised — though the structural features of the SDE system (multiplicative interactions, logistic saturation, ignition threshold, stochastic bifurcation) would survive with any alternative motivation grounding that specifies its dynamics.

Even the best formalization cannot make the recursive intelligence model empirically true. Its value is to make the theory's claims precise enough to be *clearly right or clearly wrong* — which is the only honest standard a theory can meet.

---

10. Conclusion

The Recursive Intelligence Model, as a verbal theory, identifies something that formal intelligence models systematically miss: motivation is not a confound to be controlled for but a constitutive component of the recursive dynamics that produce intellectual development. Grounding that motivation in a specific consciousness architecture — the Four-Model Theory — transforms the claim from a vague appeal to "wanting it more" into a mechanistically specified contribution of the explicit optimization layer.

The formalization recommended here preserves the verbal theory's core insights — multiplicative operational knowledge, the ignition threshold, the consciousness-grounding of motivation — while building on the correct FMT foundation: continuous model densities rather than discrete components, graded consciousness rather than binary switches, stochastic trajectories rather than deterministic curves, structured knowledge rather than scalars, and socially coupled populations rather than isolated individuals.

The result is harder. It requires tools from stochastic analysis, bifurcation theory, information geometry, and mean-field theory. But it is harder because the phenomena it describes — consciousness, intelligence, motivation, social learning — are hard. A formalization that is simpler than its subject matter is not elegant; it is wrong.

This paper specifies the program. Its execution requires mathematically trained collaborators, access to longitudinal developmental data, and the willingness to discover that some of these predictions are empirically wrong. That would be a feature, not a bug.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W. H. Freeman.

de Ron, J., Deserno, M., Robinaugh, D., Borsboom, D., & van der Maas, H. L. J. (2023). Towards a general modelling framework of resource competition in cognitive development. *Child Development*, 94(6), 1432–1453.

Driver, C. C., & Tomasik, M. J. (2023). Formalizing developmental phenomena as continuous-time dynamic systems. *Child Development*, 94(6), e336–e357.

Gruber, M. (2015). *Die Emergenz des Bewusstseins.* Self-published. ISBN 9781326652074.

Gruber, M. (2026a). Why intelligence models must include motivation: A recursive framework. *PsyArXiv* preprint. https://osf.io/preprints/osf/kctvg

Gruber, M. (2026b). 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

Gruber, M. (2026c). Toward a mathematical formalization of the four-model theory: A recommended approach. Manuscript in preparation.

Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4), 284–304.

Macnamara, B. N., & Burgoyne, A. P. (2023). Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices. *Psychological Bulletin*, 149(3–4), 133–173.

Murayama, K., & Jach, H. (2025). A critique of motivation constructs to explain higher-order behavior: We should unpack the black box. *Behavioral and Brain Sciences*, 48, e1.

Pavliotis, G. A., & Stuart, A. M. (2008). *Multiscale methods: Averaging and homogenization*. Springer.

Savi, A. O., Marsman, M., van der Maas, H. L. J., & Maris, G. K. J. (2019). The wiring of intelligence. *Perspectives on Psychological Science*, 14(6), 1034–1061.

Stamovlasis, D. (2014). Bifurcation and hysteresis effects in student performance: The signature of complexity and chaos in educational research. *Complicity*, 11(2), 51–64.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842–861.

van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review*, 98(1), 3–53.

van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos*. Harvester Wheatsheaf.