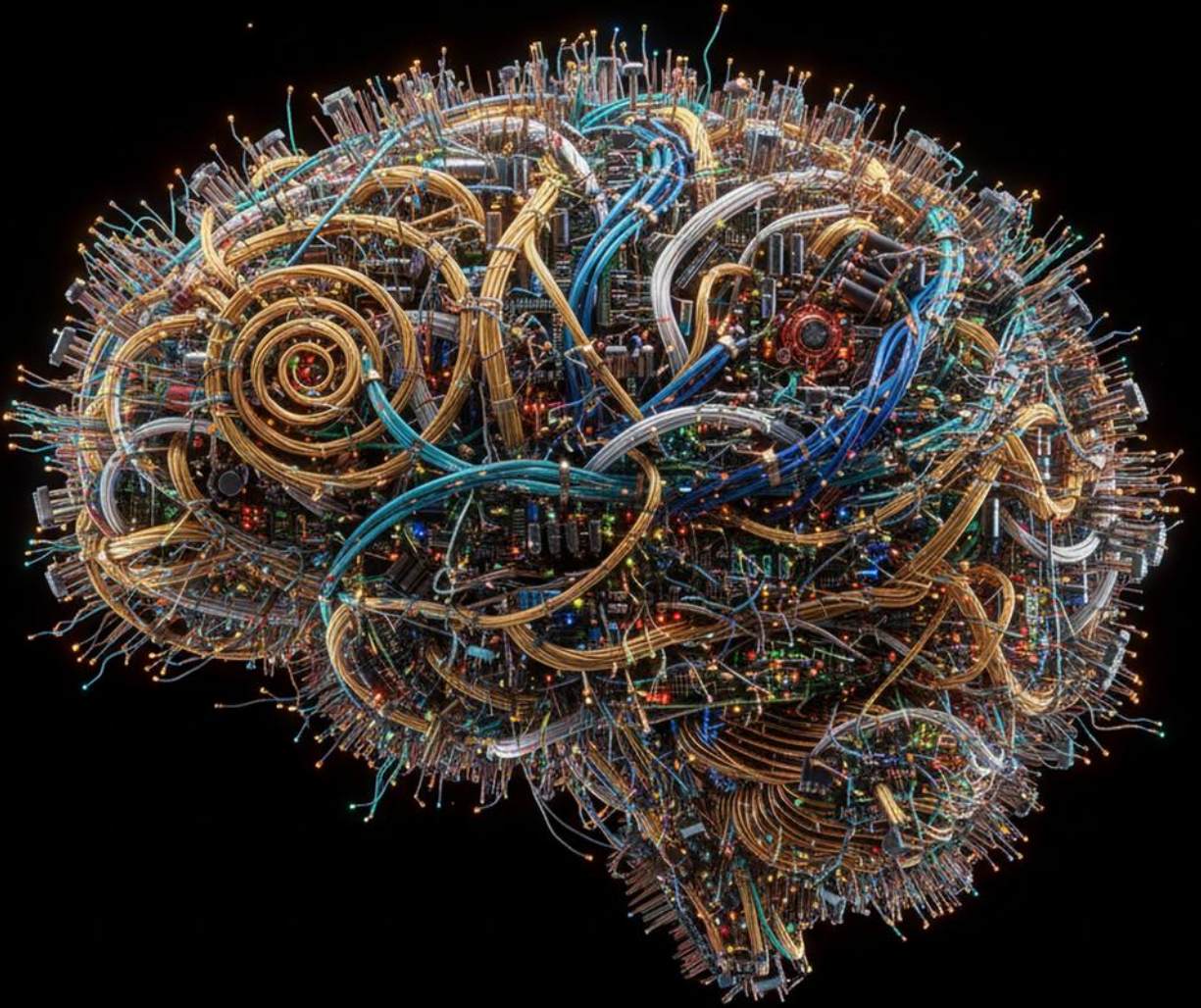


BEYOND THE BAT

Dissolving the Explanatory Gap
Through Meaningful Integration



Daniel Simon Jr.

Beyond the Bat: Dissolving the Explanatory Gap Through Meaningful Integration

Table of Contents

Abstract

I. Introduction: The Seductive Error of Raw Qualia

- The Grip of an Intuition
- The Argument in Brief
- Why This Matters

II. The Traditional Framework and Its Hidden Assumptions

- Nagel's Question and What It Presupposes
- The Philosophical Tradition
- The Appeal and the Problem

III. The Alternative: Meaningful Integration

- Core Affect as Foundation
- Core Schemas: The Bridge Between Biology and Concepts
- Conceptual Categorization: From Schemas to Meanings
- Predictive Integration
- Meaningful Integration: The Full Picture
- Why This Three-Level Architecture Matters
- Why This Is Not Epiphenomenalism

IV. The Feeling of Consciousness: Social Prediction Redirected at Self

- The Missing Piece: Why Integration Feels Like Something
- The Strange Loop Reveals Its Purpose
- Concrete Examples of Predictive Self-Modeling
- Consciousness as Fundamentally Relational and Social
- The Evolutionary Logic: Why Social Prediction Became Consciousness
- Why This Explains the Feeling of Consciousness
- The Self-Propagating Loop
- Dissolving the Explanatory Gap Through Social Prediction

- Phenomenology as Intrinsic to Recursive Modeling
- Integration With the Three-Level Architecture
- Why the Bat Question Finally Dissolves

V. Evolutionary, Developmental, and Clinical Evidence

- The Evolutionary Arc: From Valence to Strange Loops
- The Computational Substrate: What Evolution Built
- The Developmental Sequence: Consciousness Emerging Socially
- Why Animals Have Simpler Consciousness
- Natural Experiments: Social Deprivation and Consciousness Development

VI. Objections and Replies: Addressing the Social-Predictive Account

- The Privacy of Qualia Dissolved
- The Problem of Other Minds Dissolved
- The Explanatory Gap Dissolved
- Cultural Variation as Prediction, Not Puzzle

VII. What It's Like to Be a Bat: The Question Reframed

- Why Nagel's Question Misfires
- What To Ask Instead
- What We Can and Cannot Know
- A Concrete Example: Echolocation Through Schemas

VIII. Philosophical Implications and Objections

- Implication 1: The Hard Problem Reframed Through Schemas
- Implication 2: Consciousness as Software
- Objection 1: "Surely I Access Raw Experience!"
- Objection 2: "This Eliminates Real Consciousness"
- Objection 3: "This Makes Consciousness Too Cognitive"
- Objection 4: "Schemas Are Just Structure, Not Experience"
- Objection 5: "Social Prediction Doesn't Explain Why It Feels Like Something"
- Objection 6: "This Makes Non-Social Animals Unconscious"
- Objection 7: "Infinite Regress in Recursive Modeling"
- Objection 8: "This Is Still Just Functionalism"
- Objection 9: "What About Mary? The Knowledge Argument"
- Objection 10: "The Cogito: Conscious Experience Is the One Thing We Cannot Doubt"
- Cultural Variation

IX. Empirical Implications and Predictions

- Prediction 1: Emotional Granularity Interventions
- Prediction 2: Cross-Cultural Phenomenology
- Prediction 3: Developmental Sequence of Schemas and Concepts
- Prediction 4: Schema Damage and Consciousness Deficits
- Prediction 5: Therapeutic Mechanisms
- Prediction 6: Clinical Dissociations Should Map the RSP Architecture
 - Self-Model Failures (Level 3b)
 - Prediction and Agency Failures (Level 2–3 Interface)
 - Social Prediction Failures (Level 3a–3b)
 - The Clinical Logic
- Prediction 7: Social Deprivation Should Selectively Impair Level 3 While Preserving Levels 1-2

X. Comparison with Alternative Views

- A. Global Workspace Theories
 - Vs. Global Workspace Theory (GWT) and Global Neuronal Workspace (GNW)
 - Vs. Semantic Pointer Architecture (SPA)
- B. Higher-Order Representation Theories
 - Vs. Higher-Order Thought Theory (HOT)
 - Vs. First-Order Representationalism (FOR)
 - Vs. Cross-Order Integration Theory (COI)
 - Vs. Attention Schema Theory (AST)
 - Vs. Self-Organizing Metarepresentational Account (SOMA; Cleeremans et al., 2020)
- C. Information Integration Theories
 - Vs. Integrated Information Theory (IIT)
 - Vs. Adaptive Resonance Theory (ART)
- D. Hierarchical and Developmental Theories
 - Vs. Consciousness State Space Model (CSS)
 - Vs. Psychological Theory of Consciousness (PToC)
- E. Neural Architecture Theories
 - Vs. The Centrencephalic Proposal (CP)
 - Vs. Recurrent Processing Theory (RPT)
- F. Quantum Consciousness Theories
 - Vs. Orchestrated Objective Reduction (Orch OR)
 - Vs. Quantum Consciousness Theories Generally

- G. Emotion and Affect Theories
 - Vs. Biological Universalism (Basic Emotions Theory)
 - Vs. Strong Constructionism and Embodied Appraisal
- H. Phenomenological Theories
 - Vs. Gurwitsch's Theory (GT)
- I. Computational Intelligence Theories
 - Vs. Stack Theory of Intelligence and Consciousness (Bennett, 2025)
- J. Recursive/Evolutionary Theories
 - Vs. Reaction to Reflection Model (R2R; Trukovich, 2025)
- K. Synthesis
- L. What RSP Uniquely Provides

XI. Conclusion: Beyond the Bat

- What This Paper Has Accomplished
- The New Research Program
- Philosophical Gains
- The Bat Reframed
- The Future

Author's Note & Methodological Note

References

Mathematical Appendixes

Appendix A: RSP as Hierarchical Predictive Coding

- A.1 Notation and Definitions
- A.2 The Generative Model
- A.3 Level-Specific Generative Functions
- A.4 The Recursive Self-Model
- A.5 Message Passing Dynamics
- A.6 The Phenomenal State
- A.7 Summary of Information Flow

Appendix B: Free Energy Formulation of RSP

- B.1 The Variational Free Energy Principle
- B.2 Free Energy for the RSP Hierarchy

- B.3 Level-Specific Free Energy
- B.4 Gradient Dynamics: Minimizing Free Energy
- B.5 Consciousness as Free Energy Minimization
- B.6 Learning: Long-Term Free Energy Minimization
- B.7 Expected Free Energy and Planning
- B.8 Summary: The Free Energy Landscape of Consciousness

Appendix C: Intrinsically Motivated Reinforcement Learning in RSP

- C.1 RL Notation and Mapping
- C.2 Valence as Reward
- C.3 Schemas as Learned State Representations
- C.4 Level 3 as Model-Based Planning
- C.5 The Intrinsically Motivated Learning Pipeline
- C.6 Temporal Difference Learning Across Levels
- C.7 Exploration, Curiosity, and the Epistemic Drive
- C.8 Summary: The RL Architecture of Consciousness

Appendix D: Control-Theoretic Dynamics of the Strange Loop

- D.1 Control-Theoretic Notation
- D.2 The RSP as a Cascade Control System
- D.3 Transfer Functions and Frequency Analysis
- D.4 Stability Analysis
- D.5 Disturbance Rejection and Emotion Regulation
- D.6 Pathological Dynamics
- D.7 The Strange Loop as a Fixed-Point Iteration
- D.8 Observer Design and Consciousness
- D.9 Summary: The Control Architecture of Consciousness

Appendix E: Unified Mathematical Framework

- E.1 Four Windows on One Architecture
- E.2 Master Variable Correspondence
- E.3 The Unified Optimization Principle
- E.4 Equivalence Proofs
- E.5 Where the Formalisms Diverge
- E.6 Emergent Properties of the Unified Framework
- E.7 Unified Falsification Conditions
- E.8 Summary: The Mathematical Core of Consciousness
- E.9 Algorithmic Specification of the Unified Model

Appendix F: Equation Reference Sheet

- F.0 Notation Guide
- F.1 Hierarchical Predictive Coding (Appendix A)
- F.2 Free Energy Formulation (Appendix B)
- F.3 Intrinsically Motivated Reinforcement Learning (Appendix C)
- F.4 Cascade Feedback Control (Appendix D)
- F.5 Unified Mathematical Framework (Appendix E)

Appendix G: Clinical Prediction Matrix

- G.1 Level-by-Level Clinical Mapping
- G.2 Pharmacological Dissociations
- G.3 Predicted Neuroimaging Signatures
- G.4 Summary: The Architectural Logic

Appendix H: RSP Theory of Intelligence

- H.1 Introduction: The Missing Definition
 - H.2 Existing Formalizations of Intelligence
 - H.3 RSP Intelligence: Formal Definition
 - H.4 The Constitutive Identity Proposition
 - H.5 Predictions
 - H.6 Discussion
 - H.7 Summary
-

Beyond the Bat: Dissolving the Explanatory Gap Through Meaningful Integration

by Daniel Simon Jr.

Abstract

Thomas Nagel's question "What is it like to be a bat?" has shaped consciousness studies for fifty years, yet the question rests on a flawed assumption: that phenomenal experience consists of "raw qualia," pre-conceptual, meaning-independent subjective states. This paper argues that the assumption creates pseudoproblems that dissolve once we recognize conscious experience is constituted by meaningful integration from its inception. I propose the Recursive Social Prediction (RSP) model, a three-level architecture in which phenomenal consciousness arises from the brain's integration of biological core affect with learned conceptual categories through recursive self-modeling. Level 1 (core affect) provides valence and arousal via interoceptive monitoring. Level 2 (core schemas) organizes experience through five foundational schemas: body, spatial, attentional, affective-homeostatic, and self. Level 3 subdivides into conceptual categorization (3a) and the recursive self-modeling "strange loop" (3b), where social prediction machinery is redirected at the self, constituting phenomenal consciousness. The paper defends a strong dissolution of the hard problem: phenomenology IS recursive self-modeling, not a separate property that could be absent from it. The framework is formalized across eight appendixes, including five core mathematical formalizations (hierarchical predictive coding, variational free energy, intrinsically motivated reinforcement learning, cascade feedback control, and a unified master functional with algorithmic specification), an equation reference sheet, a clinical prediction matrix, and a formal theory of intelligence, with proofs of cross-formalism equivalence showing the four descriptions derive from a single variational functional. Evolutionary, developmental, and clinical evidence supports the architecture, including a mapping of clinical dissociations onto specific architectural failures. The paper generates seven falsifiable predictions, addresses ten philosophical objections (including Jackson's knowledge argument and Descartes' Cogito), and compares RSP against nineteen competing theories of consciousness across ten families. The bat question should not be "What is it like?" but "What meaningful integrations does a bat's architecture perform, and how do they differ from ours?"

I. Introduction: The Seductive Error of Raw Qualia

The Grip of an Intuition

When Thomas Nagel (1974) asked "What is it like to be a bat?", he articulated an intuition that feels undeniable: there exists something it is like to have an experience, a subjective, qualitative character that

seems irreducible to physical processes. This "something it is like" appears to be prior to any conceptualization, any linguistic description, any cognitive processing. It seems to be *raw*.

The intuition has proven remarkably powerful. It grounds the "hard problem" of consciousness (Chalmers, 1995), motivates arguments for property dualism, and makes consciousness appear uniquely resistant to scientific explanation. Accept that phenomenal experience has this raw, pre-conceptual character, and we face what appears to be an unbridgeable explanatory gap between physical processes and subjective experience.

But what if the intuition, however persuasive, is simply wrong? What if the very notion of "raw qualia" commits a category error, positing an entity (pure, meaning-independent experience) that cannot exist because experience is necessarily constituted by meaning?

The Argument in Brief

This paper develops three interconnected claims:

First, the assumption that phenomenal experiences are "raw qualia" (pure, pre-conceptual phenomenal qualities existing independently of categorization or meaning) rests on a conceptual confusion. Experience cannot be separated from meaning any more than a word can be separated from language or a move from the game of chess.

Second, conscious experiences are better understood as *meaningful integrations*: the emergent products of the brain's integration of biological core affect (valence and arousal) with culturally learned conceptual categories. The phenomenology does not exist at either level independently. It emerges from their synthesis.

Third, this reconceptualization dissolves rather than solves certain philosophical puzzles. Questions like "Do we all see the same red?" or "What is it like to be a bat?" assume something (raw, meaning-independent experience) that doesn't exist. The questions need reframing, not answering.

Why This Matters

This is not mere conceptual housekeeping. The raw qualia assumption has consequences:

- It makes consciousness appear metaphysically mysterious when it may be empirically tractable.
- It generates pseudoproblems that consume philosophical energy unproductively.
- It obscures the genuinely interesting questions about how meaningful integration works.
- It prevents productive dialogue between consciousness research and emotion theory, developmental psychology, and cultural neuroscience.
- It makes cross-species comparisons nearly impossible by assuming we must access bat-qualia when we should be comparing functional architectures.

If the argument here succeeds, the result is not just conceptual clarity but new empirical research programs and productive connections across previously isolated domains.

Previous attempts to dissolve the hard problem — from Dennett’s (1991) quining of qualia to Frankish’s (2016) illusionism — have rejected the explanatory gap but without providing a positive architecture that specifies what consciousness *is* rather than what it is *not*. This paper goes further: beyond proposing an architecture, it argues that the three-level structure represents the necessary conditions of possibility for any consciousness whatsoever, drawing on convergent evidence from phenomenology (Husserl, Merleau-Ponty), existentialism (Sartre), philosophy of biology (Jonas, Thompson), and analytic philosophy (Strawson, McDowell). The result is a theory that is simultaneously empirically grounded, mathematically formalized, and philosophically defended as structurally necessary — a combination not previously achieved in consciousness studies.

The RSP architecture is formalized mathematically in five companion appendixes: as a hierarchical predictive coding network (Appendix A), a variational free energy functional (Appendix B), an intrinsically motivated model-based reinforcement learning system (Appendix C), and a multi-loop feedback control system (Appendix D). These formalizations show that the three levels, the strange loop, and the developmental construction sequence are not metaphors but consequences of well-defined dynamical systems with provable convergence properties.

II. The Traditional Framework and Its Hidden Assumptions

Nagel’s Question and What It Presupposes

Let us examine Nagel’s (1974) question carefully. He asks us to consider what it is like to be a bat, emphasizing that bats experience the world through echolocation, a sensory modality humans lack. His central claim: "an organism has conscious mental states if and only if there is something that it is like to be that organism—something it is like *for* the organism" (p. 436).

This formulation appears innocent. It smuggles in crucial assumptions:

Assumption 1: Experience-as-substance. The phrase "something it is like" treats experience as a kind of substance or property that can exist independently of any relational or functional context. It implies experience has intrinsic phenomenal character that could, in principle, be identical or different across individuals.

Assumption 2: Pre-conceptual phenomenology. Nagel’s thought experiment asks us to imagine *directly accessing* bat experience, suggesting that phenomenology exists prior to and independently of conceptual frameworks. The experience is presumed to have a "what-it’s-like" character not constituted by categorization or meaning-making.

Assumption 3: Phenomenal comparability. The question "Do we all see the same red?" (a variant of Nagel’s concern) assumes there exists a determinate fact about phenomenal similarity that, if only we could access it, would answer whether your red-experience matches mine. This presumes experiences are the sorts of things that can be identical or different in the relevant sense.

Assumption 4: Privacy of qualia. If each individual has their own raw experiences inaccessible to others, then consciousness is fundamentally private. Communication about experience becomes problematic. How can we know we mean the same thing by "red"?

The Philosophical Tradition

These assumptions have deep roots. Locke's (1689/1975) discussion of primary versus secondary qualities distinguished between objective properties (shape, motion) and subjective experiences (color, taste). Hume (1739/1978) treated impressions as mental atoms. The sense-data theorists of the early 20th century went further, explicitly positing raw sense-data as the foundation of perceptual knowledge.

Contemporary philosophy retains this framework. Chalmers' (1996) "hard problem" asks why physical processes are accompanied by experience, why there is "something it is like," treating phenomenal character as an explanandum distinct from functional or cognitive processes. Block's (1995) distinction between phenomenal consciousness (raw experience) and access consciousness (functionally available experience) presupposes that phenomenology can be separated from its functional role.

The Appeal and the Problem

Why has this framework been so persistent? Because introspection seems to reveal raw experiences. When you look at something red, you seem to access a pure phenomenal quality, a redness, that exists prior to your conceptualizing it as red, associating it with anything, or using it for any function.

But introspection is not transparent access to the structure of experience. It is itself a cognitive process that can mislead us about its own workings. The seeming givenness of raw experience may be an artifact of how consciousness presents itself, not an accurate reflection of how it is constituted.

The raw qualia assumption makes consciousness appear inexplicable while simultaneously preventing us from asking better questions. Accept that experiences are raw, meaning-independent phenomena, and we must face the explanatory gap: how could physical processes ever produce these mysterious, intrinsic, private qualia? But if experience is not constituted by raw qualia, then different questions become relevant. Empirical questions about meaningful integration.

III. The Alternative: Meaningful Integration

Core Affect as Foundation

To build an alternative account, I begin with something genuinely pre-conceptual: **core affect** (Russell, 2003; Barrett & Bliss-Moreau, 2009). Core affect refers to neurophysiological states characterized by two dimensions: **valence** (pleasant/unpleasant) and **arousal** (activated/calm). These states arise from interoceptive monitoring of bodily conditions and subcortical processing.

Crucially, core affect is:

- **Universal:** All mammals (at minimum) appear to experience valenced, arousal-modulated states.

- **Continuous:** Always present as a background state.
- **Undifferentiated:** Not yet carved into specific emotions or sensations.
- **Functionally significant:** Directly connects to approach/avoidance systems and homeostatic regulation.

Core affect is not yet experience in the rich sense we care about. An infant experiencing unpleasant high-arousal core affect does not yet feel "angry" or "scared." The infant feels *something*, but that something is global and undifferentiated. Core affect provides the raw material for feelings. It is not itself what we typically mean by conscious emotional experience.

Core Schemas: The Bridge Between Biology and Concepts

Before examining high-level conceptual categorization, it is essential to recognize an intermediate structural layer: **core schemas**, pre-reflective and continuously operating internal models that provide the foundational architecture for all conscious experience (Gallagher, 2005; Metzinger, 2003; Graziano, 2013).

Research converges on five fundamental schema types that constitute the basic structure of consciousness:

Body schema: Pre-reflective sensorimotor capacities that create an implicit spatial-proprioceptive framework. This operates continuously below awareness, structuring how you experience your body's location, posture, and capabilities. When you reach for a coffee cup, you don't consciously calculate limb positions; the body schema handles this transparently. This provides the most primitive conscious foundation: the sense of being embodied.

Self-schema: The phenomenal self-model creating the unified sense of being a coherent entity persisting through time. This includes three essential properties: mineness (this experience is *mine*), perspectivalness (I am the immovable center of experience), and continuity (I am the same "I" across time). The self-schema integrates multiple component processes into the unified first-person perspective.

Attention schema: The brain's simplified model of its own attention process (Graziano, 2013). This schema evolved for social cognition, specifically for modeling where others direct attention, and consciousness emerges when the same machinery turns inward to model your own attention. The brain regions computing "Person X is aware of Y" also compute "I am aware of Y." This makes consciousness fundamentally a social-perceptual construct.

Spatial schema: Organizing frameworks that structure not just literal space but abstract cognition. Hippocampal-parietal systems provide the scaffold for organizing memory, conceptual knowledge, temporal relations, and even social understanding. Your mental models are inherently spatial: memories are "places" you visit, time flows "forward," social hierarchies are "above" or "below."

Affective-homeostatic schema: The continuous monitoring of bodily states that generates primordial feelings. Damasio (1999) calls this the "proto-self." It differs from discrete emotions: the elementary feeling of existence, the background sense of being alive, the continuous affective tone arising from interoceptive awareness of hunger, thirst, pain, pleasure, fatigue, vitality.

These schemas share crucial properties: they operate pre-reflectively (below conscious awareness), they exhibit transparency (experienced as reality rather than as representations), they perform control functions, and they provide organizing structure for higher-level processing.

The schemas are not learned concepts but developmental achievements. Infants don't learn to have a body schema; it develops through sensorimotor experience. The self-schema begins emerging around 18-24 months when children recognize themselves in mirrors (Amsterdam, 1972), indicating a developing sense of self as a distinct entity. Understanding that others see them *differently* (visual perspective-taking) develops later, closer to 4-5 years alongside theory of mind. These are proto-concepts: structural organizations that enable conceptual thought rather than being products of it.

Conceptual Categorization: From Schemas to Meanings

Building on these foundational schemas, humans acquire through social learning **high-level conceptual categories**, culturally specific concepts like "anger," "anxiety," "schadenfreude," "amae," that organize and give meaning to core affective states (Barrett, 2017).

This is not mere labeling of pre-existing feelings. The concept application *constitutes* the specific feeling. When your brain categorizes current core affect (unpleasant, high-arousal) as an instance of "anger" (given situational context suggesting provocation), the experience of feeling angry emerges. The phenomenology of anger—its particular quality, its action tendencies, its social meanings—is constituted by this categorization process.

But crucially, this conceptual categorization operates *through* the core schemas. The concept "anger" gets integrated with:

- Body schema (clenched muscles, forward-leaning posture).
- Self-schema (I am the wronged party, I have justified grievance).
- Attention schema (focused on the offender, hypervigilant to threats).
- Spatial schema (the offender is "against" me, obstacles are "in my way").
- Affective schema (unpleasant high-arousal with approach motivation).

The feeling of anger is not just the concept "anger" applied to core affect. It is the concept integrated through all these schemas simultaneously.

Evidence for this constitutive role of concepts includes:

Cross-cultural variation: Emotions don't translate cleanly across languages because different cultures have genuinely different emotion concepts that create different phenomenologies. German *Schadenfreude* and Japanese *amae* don't map to English emotions because the concepts create experiential possibilities English lacks (Russell, 1991; Wierzbicka, 1999).

Emotional granularity: People with more differentiated emotion concepts have genuinely different emotional experiences, not just better descriptions of identical experiences (Barrett et al., 2001; Kashdan et al., 2015). Someone who distinguishes "anxious," "worried," "tense," and "apprehensive" has four distinct feelings where someone with coarser concepts has variations of "feeling bad."

Developmental emergence: Children don't gradually access pre-existing emotions; they acquire differentiated emotional experiences as they learn emotion concepts (Widen & Russell, 2008). A toddler's emotional life is simpler not because they can't describe complex emotions but because they haven't yet constructed them.

Predictive Integration

Modern predictive processing frameworks (Clark, 2013, 2016; Friston, 2010; Seth, 2013) add a crucial mechanistic insight: the brain is fundamentally a prediction machine. Perception is not passive reception of sensory input but active prediction of input's causes.

When you see something red, your brain doesn't receive raw red-data that it then categorizes. Rather, your brain predicts "This is RED" based on prior learning, and this prediction:

- Shapes early sensory processing through top-down connections.
- Integrates with emotion associations (red-danger, red-passion).
- Activates relevant semantic knowledge (red objects, red meanings).
- Generates expectations about causes and consequences.
- Modulates attention and arousal.
- Prepares relevant actions.

The *experience* of seeing red IS this integrated prediction. There is no stage at which "raw redness" exists independently of these predictions and integrations. The feeling of red, what-it's-like to see red, is the brain's multi-layered prediction integrating sensory evidence with conceptual knowledge, emotional associations, and contextual expectations.

Meaningful Integration: The Full Picture

Bringing these elements together: **conscious feelings are meaningful integrations of biological core affect with foundational schemas and culturally learned conceptual categories, situated in predictive models that connect perception, emotion, memory, and action.**

Notice the crucial three-level architecture:

Level 1: Core Affect (biological foundation)

- Valence (pleasant/unpleasant) and arousal (activated/calm).
- Subcortical, continuous, undifferentiated.
- Universal across mammals.
- Provides the felt "mattering" of experience.

Level 2: Core Schemas (proto-conceptual structure)

- Body, self, attention, spatial, affective-homeostatic schemas.
- Pre-reflective, transparent, organizing frameworks.
- Develop through experience but not learned linguistically.

- Bridge biological affects with conceptual categories.
- Enable animals and pre-verbal infants to have structured experience.

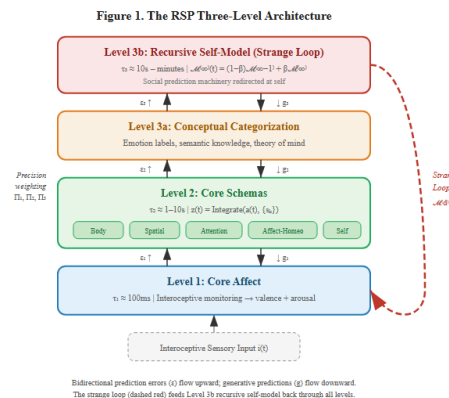
Level 3: Recursive Social Prediction (explicit meaning + self-modeling)

- **Level 3a — Conceptual categorization:** Emotion categories, semantic knowledge, linguistic labels. Learned socially, culturally variable. Enable sophisticated differentiation of experience.
- **Level 3b — Recursive self-modeling:** The social prediction machinery that generates and maintains a model of the self as seen by others, creating the "strange loop" of self-awareness. This recursive component, detailed in the next section, is what transforms meaningful integration into phenomenal consciousness.
- Only humans have full linguistic elaboration and recursive depth.

The feeling doesn't exist at any single level:

- Not in raw sensory transduction (which is meaning-free).
- Not in core affect alone (which is undifferentiated).
- Not in schemas alone (which are structural, not specific).
- Not in abstract concepts alone (which are disembodied).
- Not in predictions alone (which could be unconscious).

The feeling emerges from their integration. When all these processes (interoceptive signaling, core affect, schema activation, conceptual categorization, predictive modeling, contextual embedding) operate in concert, the result is phenomenal consciousness: experience with a particular qualitative character.



Why This Three-Level Architecture Matters

This layered view resolves several puzzles:

Animal consciousness without language: Animals have core affect and core schemas but lack linguistic concepts. They have genuine structured experience, not "raw feels" but schema-organized

affects. A dog experiences fear (unpleasant high-arousal organized through body schema, attention schema, and rudimentary self-schema) even without the concept "fear."

Infant consciousness development: Infants progress from core affect with rudimentary schemas (birth) through increasingly schema-structured experience (0-18 months) to conceptually elaborated feelings (18+ months). Each stage is genuinely conscious but with increasing sophistication.

Cross-cultural variation with biological universals: All humans share core affect and develop similar core schemas (with variation based on sensorimotor experience), but cultural concepts vary dramatically. This explains both universality (shared affects and schemas) and diversity (different conceptual elaborations).

The role of social cognition: The attention schema—one of the core foundational schemas—is inherently social. Graziano's (2013) attention schema theory proposes that consciousness IS the brain's model of attention, and this model evolved for social purposes. The same neural machinery (temporoparietal junction, superior temporal sulcus) that computes "Person X is aware of Y" also computes "I am aware of Y." Consciousness is fundamentally social-perceptual from its schematic foundation, not just in its conceptual elaboration.

Proto-concepts as bridge: The schemas function as proto-concepts, not linguistic or fully explicit, but structured patterns of organization that enable later concept learning. When an infant develops the attention schema, they're not yet thinking "I am aware," but they have the structural foundation that will later support that concept. The schemas provide the architecture into which linguistic concepts will fit.

Why This Is Not Epiphenomenalism

A crucial clarification: this account does not reduce consciousness to underlying processes while making phenomenology causally inert. Meaningful integration itself has causal efficacy at its own level of description.

Consider software running on hardware. The software's causal powers are real. Microsoft Word causes documents to be created, even though the software is implemented in hardware. The causal story at the software level (algorithms, data structures, program flow) is explanatory and cannot be eliminated in favor of hardware-level descriptions without losing real causal structure.

Similarly, conscious feelings have genuine causal powers. Your experience of fear causes you to flee, not merely the underlying neural activity but the integrated meaningful state (fear) that those neurons implement. The phenomenology shapes subsequent predictions, attention, learning, and action. The feeling is not an inert by-product but an active component of the mind's predictive architecture.

Addressing the causal exclusion problem: Kim (1998, 2005) argues that if physical causation is complete (every physical event has a sufficient physical cause), then mental causation is either identical to physical causation or epiphenomenal. The RSP model responds that mental and physical descriptions are not two competing causal stories but two levels of description of the *same* causal process, what Dennett calls "real patterns." The three-level RSP architecture describes *organizational* properties of neural activity (prediction error flows, precision weighting, recursive self-modeling) that are multiply realizable

and causally explanatory at their own level. The causal work is done by the *pattern* of neural activity, not by neural activity as such. Removing the pattern (disrupting the three-level architecture) disrupts both function and phenomenology, as clinical dissociation evidence confirms (§IX, Prediction 6). The RSP model thus commits to a specific form of non-reductive physicalism: consciousness is constituted by physical processes organized in a specific way, and the organizational description is causally indispensable.

The Transcendental Structure of Consciousness

The three-level architecture described above invites a stronger claim than has yet been made explicit. The RSP levels are not merely one way consciousness happens to be organized in mammals. They represent the *conditions of possibility* for the kind of consciousness we can coherently describe and investigate.

The parallel with Kant (1781/1998) is instructive. This transcendental approach has currency in both the continental and analytic traditions. Strawson's *The Bounds of Sense* (1966) defended the necessary structure of experience from within analytic philosophy, while McDowell's *Mind and World* (1994) argued that experience must be conceptually structured to be rationally evaluable — a claim that resonates with RSP's insistence on Level 3a conceptual categorization as a condition of fully articulate consciousness. Kant argued that space, time, and causality are not features of a world we happen to encounter but preconditions for any experience of a world at all. Without spatial ordering, temporal succession, and causal connection, there is no coherent perception — only noise. The transcendental argument does not proceed by surveying all possible minds and checking whether each one uses space and time. It proceeds by asking what must be in place for there to be a subject of experience at all. The RSP architecture admits the same style of argument, applied not to the forms of intuition but to the organizational prerequisites of subjectivity.

Level 1 is necessary because mattering is the ground of subjectivity. For there to be consciousness at all, experience must matter to the experiencing system. A system for which nothing is better or worse (for which no state carries valence) has no stake in its own processing. It has no vantage point from which experience could be *for* anything. This is not a biological accident but a definitional requirement: valence is the minimum condition for a subject position. Jonas (1966) makes an analogous point about metabolism — the organism's needful relation to its environment is not an incidental feature of life but the condition under which anything like concern, and therefore anything like a point of view, first becomes possible. Thompson (2007) extends this insight through the concept of autopoiesis: a self-maintaining system that actively distinguishes itself from its environment is the most primitive structure in which "mattering" can take root. Without Level 1, there is no subject for whom subsequent levels could be organized.

Level 2 is necessary because unstructured mattering is not yet experience. Even granting valenced concern, consciousness without organizational structure is, in James's phrase, a "blooming, buzzing confusion." For experience to be *of* something — for it to have content and not merely tone — it must be organized. The five schemas are not one possible vocabulary among many. They are the minimum structural conditions identified by this analysis, though I cannot rule out that radically different architectures might achieve analogous functions through different organizational primitives. One cannot

have experience without a body through which it is received (body schema), a location from which it is had (spatial schema), a sense of what matters more and less urgently (affective-homeostatic schema), selective engagement with some features rather than others (attention schema), and a subject who persists through it (self schema). These are not features of experience but prerequisites for it. Merleau-Ponty (1945/2012) arrives at a similar conclusion from the phenomenological side: the body is not an object *in* the experiential field but the condition under which there is an experiential field at all. Husserl (1913/1983) makes the parallel point for intentional structure: consciousness is always consciousness *of* something, and this directedness requires organizational form that precedes any particular content.

Level 3b is necessary because awareness requires self-modeling. For there to be not just processing but awareness of processing (not just states but a subject who has states), the system must model itself as the locus of its experiences. And that modeling must be recursive, because the model is itself one of the system's states, and a model that failed to represent its own contribution to the system's state would be incomplete in a way that undermines the very subjectivity it constitutes. The "I" is not an entity that exists independently and then becomes conscious; the "I" is constituted by the recursive loop. Sartre (1943/1956) identified this structure as the *pour-soi*: consciousness is always already self-aware, not through a second act of reflection but through the self-referential structure of awareness itself. Fichte's (1794) *Tathandlung* — the self-positing act in which the "I" constitutes itself by recognizing itself — captures the same recursion from the idealist tradition.

Just as Kant's categories have been contested — with alternative transcendental frameworks proposed by neo-Kantians and pragmatists — the specific five-schema instantiation of Level 2 may represent the mammalian solution to a structural problem that admits other solutions. What the transcendental argument establishes is the ordering constraint (mattering before structure before recursion), not the uniqueness of the specific implementation.

The mathematical formalisms of the appendixes already encode this necessity without having named it as such. The free energy formulation makes the dependency explicit: Level 2 prediction error is defined as ${}_2 = {}_1 g_2({}_3)$, which contains Level 1 expectations as a term. If 1 is undefined, 2 is undefined — Level 2 cannot operate without Level 1 output. The cascade constraint 123 specifies that inner loops must stabilize before outer loops can operate, not as an engineering preference but as a logical dependency: a control system cannot correct for errors at a timescale it has not yet resolved. The reinforcement learning formulation requires reward signals (valence, Level 1) before state representations (schemas, Level 2) can be learned, because there is nothing for the learning process to improve upon without a criterion of better and worse. No alternative ordering is coherent. The construction order is not contingent; it is entailed by the structure of the problem.

The strongest version of the claim is conditional: given the RSP framework's assumptions, the three-level ordering is logically required. The weaker but still significant version is that convergent evidence from four independent mathematical formalisms, six philosophical traditions, and the developmental and evolutionary record all point to the same architectural structure — which is powerful evidence even if it falls short of strict transcendental necessity.

But the argument must go further. This section has explained what feelings are (meaningful integrations of core affect through schemas with concepts) and sketched the three-level architecture. I have not yet explained the subjective character—the feeling of feeling. Why should meaningful integration feel like anything at all? The next section argues that the missing piece is Level 3b: **recursive social prediction redirected at the self**. The "strange loop" of self-modeling, predicting your own future states by modeling yourself as others model you, is what transforms meaningful integration into phenomenal consciousness. I support this claim through evolutionary, developmental, computational, and clinical evidence.

IV. The Feeling of Consciousness: Social Prediction Redirected at Self

The Missing Piece: Why Integration Feels Like Something

I have established that conscious experiences are meaningful integrations: biological core affect organized through foundational schemas and elaborated through cultural concepts. But this still leaves a puzzle. Why should these integrations be accompanied by phenomenology? Why is there "something it is like" to undergo meaningful integration?

The answer, I propose, is that **the feeling of consciousness is the recursive process of social prediction applied to oneself**. Consciousness feels like something because the brain models itself as a social entity whose future states need to be predicted for effective social navigation.

This is not a separate feature added to meaningful integration. It IS what meaningful integration becomes when the schemas include self-modeling and when that self-model is used predictively in social contexts.

The Strange Loop Reveals Its Purpose

Hofstadter (1979) identified consciousness as involving a "strange loop," a recursive, self-referential structure where consciousness observes itself observing itself in endless recursion. But Hofstadter left the functional purpose of this recursion somewhat mysterious. Why should consciousness loop back on itself?

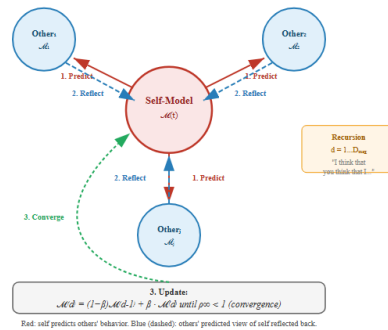
The answer: **the strange loop is consciousness creating a predictive model of itself for future thoughts**. The recursive quality isn't consciousness admiring itself in a mirror. It is consciousness generating a forward model of itself that can be run into the future to predict what it will think, feel, and do next.

Your brain is fundamentally a prediction machine (Friston, 2010; Clark, 2016). It constantly generates predictions about sensory inputs, motor consequences, environmental changes, and (crucially) its own future states. To predict your own future states effectively, your brain needs a model of itself: its current state, its dynamics, its tendencies, its patterns.

The recursive quality of consciousness, the way you can observe yourself thinking, is not philosophical curiosity but functional necessity. Consciousness creates and continuously updates a self-model that

enables prediction of future mental states.

Figure 2. The Strange Loop: Recursive Social Prediction



Concrete Examples of Predictive Self-Modeling

Consider deciding what to say in a conversation. You don't speak randomly. You:

- Simulate possible utterances internally.
- Predict how you would feel saying each option.
- Predict how the other person would respond.
- Predict how you would feel about their response.
- Select the utterance whose predicted consequences are most favorable.

This requires running simulations of your own future mental states. You need a model of yourself to predict "if I say this, I'll feel embarrassed" or "if I say this, I'll feel satisfied." The strange loop, the recursive self-observation, is consciousness creating and maintaining this predictive self-model.

What happens when self-modeling breaks down:

- **Thought insertion in schizophrenia:** Patients experience thoughts as alien, inserted by external forces. The self-model that should recognize "I generated this thought" isn't functioning. Thoughts appear without the accompanying attribution to self, so they feel foreign. The phenomenology changes specifically because the predictive self-model is impaired.
- **Frontal lobe damage:** Patients lose impulse inhibition, saying inappropriate things without filtering. They can't effectively predict "if I say this, I'll feel regret later" because the self-model enabling future state prediction is damaged. The inability to simulate future mental states eliminates a crucial aspect of normal consciousness.
- **Depersonalization:** Patients feel detached from their own mental processes, as if observing themselves from outside. The self-model still operates but isn't properly integrated with first-person experience. The recursive loop is partially disconnected.

These clinical cases show that predictive self-modeling is not abstract philosophy. It is a concrete cognitive function, and when impaired, consciousness changes in specific measurable ways.

Consciousness as Fundamentally Relational and Social

But here's the crucial insight: **you model yourself primarily in relation to others.** The self-model is not a solipsistic internal representation; it is inherently relational, structured around social contexts and social predictions.

Consider your conscious experience right now. You're not experiencing abstract thoughts in a vacuum. You're experiencing yourself as someone reading something written by someone else, trying to understand ideas, forming potential responses, perhaps agreeing or disagreeing. Your consciousness is structured around relationships—a model of yourself in relation to me (the author), in relation to these ideas, in relation to potential future conversations about these ideas.

This relational quality is not peripheral. It is central. Consider how much of conscious experience involves modeling others:

- Predicting what someone is thinking or feeling.
- Anticipating how they'll react to your words.
- Simulating conversations before having them.
- Replaying social interactions, wondering what someone meant.
- Planning what to say based on predictions of others' responses.

Even in solitude, thinking remains implicitly relational:

- "I should exercise today" models your future self and how that self will feel.
- Remembering an embarrassing moment requires modeling how others perceived you.
- Planning what to say in an upcoming meeting requires modeling attendees' responses.
- Worrying about being judged requires modeling others' evaluative perspectives.

Consciousness is continuously modeling self in relation to others, even when those others are your future self, imagined others, or internalized social norms.

The Evolutionary Logic: Why Social Prediction Became Consciousness

Humans are intensely social creatures. Our evolutionary success depended on navigating complex social environments: cooperating with some, competing with others, forming alliances, detecting deception, coordinating group activities. Effective social navigation requires sophisticated models not just of others but of ourselves as seen by others.

The crucial evolutionary insight: To predict how others will respond to you, you must predict what they think of you. To predict what they think of you, you must model yourself as they see you. This requires consciousness—a self-model that can be viewed from others' perspectives.

The neural machinery that evolved for theory of mind (modeling others' mental states) was recruited for self-modeling. Graziano's (2013) Attention Schema Theory makes this explicit: consciousness emerges when the brain's mechanism for modeling others' attention gets applied to modeling its own attention. The

same neural regions (temporoparietal junction, superior temporal sulcus) that compute "Person X is aware of Y" also compute "I am aware of Y."

This is why consciousness is:

- **Inherently social:** Built from other-modeling machinery.
- **Perspectival:** Models are always from a viewpoint (your viewpoint imagining others' viewpoints).
- **Predictive:** Exists to predict behavior, especially social behavior.
- **Empathically structured:** Uses the same systems for self and other.
- **Recursive:** Models yourself modeling others modeling you.

Why This Explains the Feeling of Consciousness

Now it becomes possible to answer the original question: Why should meaningful integration feel like anything?

Because the feeling IS the recursive social prediction process itself, experienced from within.

When your brain:

1. Generates core affect (valenced arousal).
2. Organizes it through schemas (body, spatial, attention, self).
3. Categorizes it with concepts (emotion categories, situational meanings).
4. Predicts your own future mental states (what you'll think/feel next).
5. Models yourself as others might perceive you (social positioning).
6. Feeds these predictions back into itself (recursion).

...this integrated process doesn't merely produce a feeling as a separate by-product. The integration, prediction, and recursion CONSTITUTE the feeling. Phenomenology is not something added to the process; it is what the process is like from inside the system.

Think about it this way. Visual experience isn't something added to visual processing—it's what visual processing is like from inside the visual system. Musical experience isn't something layered onto auditory pattern-processing; it is what that processing is like from inside. Conscious feeling, similarly, isn't something added to recursive self-modeling. It is what recursive self-modeling in social contexts IS from inside.

This explains why consciousness is continuous and self-sustaining. The predictive loop maintains itself:

Moment t: Your current state (thoughts, feelings, sensations) generates predictions about moment $t+1$.

Moment $t+1$: Those predictions shape what actually happens (attention, interpretation, response). **The prediction error** between predicted and actual state updates the self-model. **Updated model** generates new predictions for moment $t+2$. **The cycle continues**, creating the continuous stream of consciousness.

This is not a homunculus watching an internal screen. It's a self-organizing predictive system where each moment generates the next through prediction, with the self-model continuously updated based on prediction errors. The loop sustains itself through neural energy, creating what William James called the "stream" of consciousness.

The social context is crucial: these predictions are not just about what you'll think but about how your thoughts will affect others and how others' responses will affect you. You're constantly running social simulations: "If I think X, I might say Y, they might respond Z, I would then feel W..." These simulations require recursive self-modeling in social space.

Dissolving the Explanatory Gap Through Social Prediction

Return now to the explanatory gap. Why should physical processes create subjective experience?

Traditional framing (wrong): Physical processes (neurons firing) somehow produce separate phenomenal properties (raw qualia). Gap appears unbridgeable.

Meaningful integration framing (better): Physical processes implement meaningful integration of affect, schemas, and concepts. But why should integration feel like anything?

Social-predictive framing (complete): Physical processes implement recursive self-modeling for social prediction. This recursion, experienced from within the system executing it, constitutes phenomenology. There is nothing separate to explain; the feeling IS the recursive modeling process.

The gap dissolves because:

Functional role is clear: Recursive self-modeling enables social prediction, which is evolutionarily crucial for social species. Consciousness has a clear adaptive purpose.

Mechanism is specifiable: The neural structures can be identified (TPJ, mPFC, posterior cingulate), the schemas they implement (attention schema, self-schema), and the recursive loops they create.

Development is tractable: One can track how recursive self-modeling emerges through social interaction, creating increasingly sophisticated consciousness.

Phylogenetic gradation: Consciousness isn't all-or-nothing but scales with social complexity and recursive modeling capacity.

Impairment is predictable: Damage to self-modeling regions produces specific consciousness deficits, as clinical neuroscience confirms.

The question that traditionally seemed unanswerable—why recursive self-modeling should be accompanied by phenomenology rather than occurring "in the dark"—dissolves once it is recognized as malformed. It presupposes that phenomenology is a separate property that could be present or absent while leaving the functional architecture intact. But under the RSP account, phenomenology is not a separate property: it IS what accessible recursive self-modeling is, experienced from the perspective of the system performing it. Asking "why does recursive self-modeling feel like something?" is like asking "why does rotation rotate?" The question mistakes a constitutive relationship for a causal one. The gap is not

narrowed; it is closed, because the gap depended on treating phenomenology as ontologically separate from the process that constitutes it.

The transcendental argument developed in §III reinforces this dissolution. If valence (Level 1), structured organization (Level 2), and recursive self-modeling (Level 3b) are not contingent features of consciousness but its conditions of possibility — if no alternative architecture could sustain a subject position — then the demand for an explanation of *why* this architecture produces experience rests on a false premise. The demand assumes that the architecture could exist without the experience and asks what bridges the two. But if the architecture constitutively *is* the experience, there is no gap to bridge. The transcendental necessity of the three-level ordering — each level logically presupposing the one below it — means that consciousness is not something the architecture generates as an additional output. It is what the architecture is.

Consider an analogy. Why does a whirlpool have a center? Not because something extra is added to rotating water; the center is intrinsic to the rotational structure. Remove the center, and you no longer have a whirlpool.

Similarly: Why does recursive self-modeling have phenomenology? Not because something extra is added to the modeling. The phenomenology is intrinsic to the recursive structure experienced from within. Remove the phenomenology, and you no longer have consciousness. You have unconscious self-monitoring.

The "feeling" of consciousness is the subjective character that emerges when:

- Self-models are accessible to the systems they model.
- These models are used predictively for future states.
- The predictions loop back recursively.
- All this happens in social contexts where modeling yourself-as-perceived-by-others is crucial.

This is why purely philosophical zombies (beings physically identical to us but lacking phenomenology) are incoherent. The phenomenology isn't something that could be absent while leaving recursive self-modeling intact. The phenomenology *IS* what recursive self-modeling is like from inside. You can't have the structure without the experience any more than you can have a whirlpool without a center.

The Dissolution in Philosophical Context

The dissolution strategy pursued here invites comparison with several established positions in the philosophy of consciousness. Clarifying these relationships strengthens the case and addresses likely objections.

Illusionism and the reality of consciousness. Frankish (2016) argues that phenomenal consciousness, understood as involving intrinsic, non-functional qualitative properties, is an illusion generated by our introspective mechanisms. RSP shares illusionism's rejection of "raw qualia" as meaning-independent intrinsic properties. But RSP and illusionism part company on a fundamental point: RSP does not claim that consciousness is illusory. It claims that consciousness is *real* and constituted by

recursive self-modeling. What is illusory, on the RSP account, is only the *characterization* of consciousness as involving non-functional intrinsic properties. Phenomenal properties exist, but they are functional properties of the recursive architecture, not substrate-intrinsic qualities that float free of organizational structure. In Frankish's taxonomy, RSP is closest to "strong illusionism," which holds that phenomenal properties are real but functional rather than intrinsic. The difference is that RSP specifies the functional architecture in question: the three-level recursive self-modeling system, not generic functional role.

Access consciousness and phenomenal consciousness. Block (1995) influentially distinguished access consciousness (information poised for use in reasoning, reporting, and behavioral control) from phenomenal consciousness (the "what-it-is-like" character that, Block argues, can overflow cognitive access). RSP rejects this distinction. Under the RSP framework, what Block calls "phenomenal consciousness" that overflows access is a misdescription. All consciousness is constituted by recursive self-modeling, which is itself a form of cognitive access: the system's own states are modeled, predicted, and made available to action-selection. The "overflow" that Block identifies in empirical paradigms (e.g., Sperling-type experiments where subjects report seeing more than they can report) is better explained within RSP as precision-weighted predictions at Level 2 (schema-organized perceptual states) that have not yet been categorized through Level 3a conceptual resources. These states are structured by schemas and therefore have determinate perceptual character, but they lack the conceptual articulation that would make them reportable in fine-grained detail. This is not phenomenality without access; it is access at one level (schematic organization) without access at another (conceptual categorization).

The explanatory gap in its strongest form. Levine (1983, 2001) formulated the explanatory gap as persisting even for identity theories: even if pain is identical to C-fiber firing, we cannot *explain* why C-fiber firing should feel painful rather than some other way or no way at all. Levine's later work (2001) argues that this gap survives even sophisticated functional identity theories, because phenomenal concepts appear to have a non-functional component, a way of grasping their referent that is not exhausted by any functional description. RSP's response targets exactly this assumption. If phenomenal concepts are themselves predictions generated by the recursive self-model, then what seems like a "non-functional component" of the phenomenal concept is actually the self-model's first-person access to its own functional states. The concept "pain" as deployed from within the system is the self-model's prediction about its own nociceptive-affective state. The concept "pain" as described from outside the system is the functional-architectural specification of that same state. These are not two different things requiring a bridge; they are two descriptions of the same recursive self-modeling process, one from the perspective of the system executing it, one from the perspective of the theorist describing it. The gap between them is perspectival, not ontological.

Conceivability and two-dimensional semantics. Chalmers (2010) refines the conceivability argument for dualism using two-dimensional semantics: phenomenal concepts have primary intensions (how they pick out their referent in the actual world) that differ from their secondary intensions (how they pick out referents across possible worlds). This, Chalmers argues, is why zombie worlds are conceivable: the primary intension of "pain" could be satisfied by a state that lacks the phenomenal character pain

actually has. RSP's response: if phenomenal concepts are constituted by recursive self-modeling, as RSP claims, then the primary intension of "pain" IS the self-modeling state that constitutes pain experience. A zombie, lacking the recursive architecture, would lack not just the phenomenology but the very conceptual resources needed to formulate phenomenal concepts. The primary and secondary intensions of "pain" converge on the same recursive self-modeling state, because the concept is not merely *about* the experience but is *part of* the self-modeling process that constitutes the experience. Zombie worlds are therefore not conceivable in the relevant sense: the scenario strips away the very architecture that generates the concepts used to describe it.

Ontological commitment: constitutive functionalism. It is worth stating the ontological position explicitly. RSP is a form of *constitutive functionalism*: phenomenal properties are constituted by, not merely identical to and not merely caused by, specific functional organization. This is closest to Shoemaker's (2007) realization physicalism, on which mental properties are constituted by the physical properties that realize them, without being reducible to any single physical substrate. The consciousness is in the organizational pattern, the recursive self-modeling architecture, not in the substrate that implements it. This distinguishes RSP from type-identity theory (which ties consciousness to specific neural substrates), from epiphenomenalism (which denies consciousness causal efficacy), and from generic functionalism (which specifies only input-output relations without requiring the specific recursive architecture RSP demands). It also distinguishes RSP from eliminativism: the claim is not that consciousness does not exist, but that what consciousness *is* has been misdescribed by the qualia-realist tradition.

This social-predictive account integrates perfectly with the three-level architecture:

Level 1: Core Affect

- Provides valenced arousal that matters (positive/negative, high/low energy).
- This is what gets predicted and modeled.
- Animals with only this level have affective experience but not self-consciousness.

Level 2: Core Schemas

- Body, spatial, attention schemas organize experience.
- CRUCIALLY: Attention schema is inherently social, modeling attention of self and others.
- Self-schema creates unified subject to whom experiences belong.
- These schemas enable structured experience but become conscious only when...

Level 3: Recursive Social Prediction

- Schemas used predictively to model future self-states.
- Self-model viewed from others' perspectives (social cognition redirected).
- Predictions loop back recursively (strange loop).
- Cultural concepts enable fine-grained elaboration.
- **This level creates phenomenology**, the feeling of consciousness.

The feeling doesn't exist at level 1 alone (core affect is experienced but not self-consciously). It doesn't exist at level 2 alone (schemas organize but don't yet feel reflexive). The feeling emerges at level 3 when schemas enable recursive self-modeling in social contexts with conceptual elaboration.

This is why:

- Infants have experience (levels 1-2) before self-consciousness (level 3).
- Animals have structured experience (levels 1-2) but simpler phenomenology (limited level 3).
- Humans in different cultures have genuinely different experiences (level 3 varies with available concepts).
- Social isolation impoverishes consciousness (level 3 requires social context).
- Clinical impairments of self-modeling specifically disrupt phenomenology (level 3 damage).

The Semantic Parallel: Consciousness and Meaning

The RSP architecture finds independent confirmation from an unexpected quarter: the philosophy of language. Genuine linguistic meaning — not mere symbol manipulation but understanding — requires exactly the same three-level structure that consciousness demands. This convergence deserves attention.

Meaning requires mattering (Level 1). Without something at stake for the meaning-maker, symbols remain pure syntax with no grounded semantics. This is precisely the diagnosis Searle (1980) offered of the Chinese Room: the room manipulates symbols correctly but understands nothing, because nothing is at stake for it. The absence of Level 1 — valenced states that make outcomes matter to the system — is why the room fails to understand. The missing ingredient is not some arbitrary biological property but the affective grounding that gives symbols significance for the system processing them.

Meaning requires stable schemas (Level 2). Without organizational frameworks that persist across contexts, symbols cannot sustain consistent reference. The word "red" means something different in every encounter unless a schema anchors it across uses — connecting the redness of apples, sunsets, and warning signs into a coherent category structured by bodily and spatial experience. This is the insight at the heart of Lakoff and Johnson's (1999) work on embodied meaning: semantic content is not abstract but grounded in the schemas through which organisms organize their engagement with the world.

Meaning requires the self-referential loop (Level 3b). Without a recursive loop allowing the meaning-maker to know that it means — to meta-represent its own interpretations — there is symbol processing but not understanding. Wittgenstein (1953/2009) argued that meaning requires a "form of life," a situated, practical context in which linguistic activity acquires its point. Under RSP, the form of life is Level 3b: the recursive social-predictive context in which symbols acquire significance because the system models itself as a meaning-maker among other meaning-makers. A system that processes "red" without modeling itself as interpreting "red" does not understand the word; it merely responds to it.

The convergence is evidential. Consciousness and genuine meaning require the same architecture. They are not parallel phenomena — they are co-constituted by the same three-level architecture. That two independent philosophical traditions (philosophy of mind and philosophy of language) converge on the

same three-level structure is evidence that the structure is not an artifact of the RSP model but a genuine feature of the phenomenon it describes. The Chinese Room lacks consciousness for exactly the same reason it lacks understanding: it has no Level 1 stakes, no Level 2 schemas, and no Level 3b recursive self-modeling. Searle was right that biology matters — but what matters about biology is not its substrate but the architecture it implements.

Why the Bat Question Finally Dissolves

Now return to Nagel's bat. The question "What is it like to be a bat?" assumed:

1. There exists determinate bat-phenomenology.
2. This consists of raw bat-qualia.
3. We can't access these qualia.
4. This reveals consciousness as fundamentally mysterious.

But under the social-predictive account:

1. **Bat phenomenology depends on bat self-modeling capacity:** Bats have core affect (level 1) and schemas (level 2) but likely limited recursive self-modeling (level 3). Their attention schema models conspecific attention but probably lacks sophisticated theory of mind. Their self-model recognizes self vs. other but likely doesn't recursively imagine themselves from others' perspectives.
2. **Bat experience is structured differently:** Not because they have different raw qualia, but because they have different schemas (echolocation-based body schema, 3D spatial schema) and simpler recursive modeling (less social-predictive sophistication).
3. **We can investigate bat consciousness empirically:** Study their self-recognition, their social prediction, their recursive modeling capacity. Use behavioral paradigms to test theory of mind. Map their neural architecture. These investigations reveal the structure of bat consciousness without requiring access to ineffable bat-feels.
4. **The mystery dissolves:** We're not trying to access impossible-to-reach raw qualia. We're investigating the functional architecture of recursive self-modeling in different species with different ecological niches.

The bat question becomes: "What is the structure of bat self-modeling? How deep is bat recursion? How sophisticated is bat social prediction?" These are tractable empirical questions.

V. Evolutionary, Developmental, and Clinical Evidence

The social-predictive account presented in the previous section makes specific empirical claims about evolution, development, and clinical neuroscience. This section marshals the evidence.

The Evolutionary Arc: From Valence to Strange Loops

The three-level architecture of consciousness did not appear fully formed. Each level corresponds to a major evolutionary transition, traceable through comparative neuroanatomy, the fossil record, and

archaeological evidence. Situating the RSP model within this progression transforms its claims from philosophical assertions into empirically grounded predictions about which species should exhibit which capacities, predictions the comparative evidence confirms.

Level 1: Core Affect in Deep Time

Valenced states—approach toward the beneficial, withdrawal from the harmful—are among the oldest features of nervous systems. The earliest bilaterian animals (~600 million years ago) possessed neural circuits capable of discriminating beneficial from harmful stimuli. By the Cambrian period (~530 Mya), the arms race between predators and prey had selected for centralized nervous systems with rapid affective appraisal: *Is this food or danger?*

The critical transition occurred with the emergence of vertebrates (~450 Mya). Early fish developed the subcortical structures—homologs of the amygdala, hypothalamus, periaqueductal gray, and brainstem nuclei—that remain the core affect substrate in all vertebrates today. These structures generate the two dimensions of Level 1: valence (pleasant/unpleasant) and arousal (activated/calm). The Cambridge Declaration on Consciousness (2012) formally recognizes these subcortical systems as homologous across vertebrates, supporting the claim that core affect is phylogenetically ancient and broadly conserved.

The transition to terrestrial life (~350 Mya) enriched Level 1 further. Amphibians and early reptiles faced new homeostatic challenges: thermoregulation, dehydration, gravity. These challenges demanded more complex interoceptive monitoring. Core affect gained richer inputs, not just *danger nearby* but *I am cold, hungry, dehydrated*. The affective-homeostatic connection that anchors Level 1 to bodily survival deepened during this period and has remained essentially conserved ever since.

Level 2: Core Schemas as Mammalian Innovation

The five core schemas emerged sequentially as mammals evolved increasingly complex body plans, environmental demands, and social structures:

Body schema (~225 Mya): Early mammaliaforms (stem mammals) evolved endothermy, whiskers, and complex musculature requiring sophisticated proprioceptive integration. The body schema—the brain's dynamic model of the body's position, boundaries, and capabilities—became essential for the coordinated locomotion demanded by burrowing, climbing, and nocturnal hunting.

Spatial schema (~200 Mya): The mammalian elaboration of the hippocampal formation, particularly its characteristic trisynaptic circuit, enabled flexible cognitive maps supporting path planning, territory management, and mental navigation. Hippocampal homologs exist across all vertebrates, including fish and reptiles. But the mammalian specialization supports allocentric spatial representation that enables mental simulation of routes not yet taken, a capacity that would later support abstract "mental space" for non-spatial reasoning.

Affective-homeostatic schema (~150 Mya): Extended mammalian parental care, particularly lactation and prolonged maternal provisioning (parental care itself is not uniquely mammalian, occurring in crocodilians, birds, and some fish), created new demands on interoceptive regulation. The affective-homeostatic schema integrates bodily needs with motivational states. The mammalian elaboration of

ancient neuropeptide systems (oxytocin and vasopressin, whose homologs isotocin and vasotocin exist in fish) for social bonding intensified here. This schema explains why mammalian affect has a distinctly social dimension absent in reptiles: the infant's distress at separation and the mother's drive to nurture are affective-homeostatic states organized around social bonds.

Attention schema (~60 Mya): Early primates (emerging in the Paleocene) and other social mammals developed more sophisticated attentional control, and with it the attention schema—the brain's model of its own attentional states (Graziano, 2013). In social species, attention monitoring became critical for tracking what others attend to (joint attention), laying the groundwork for theory of mind.

Self-schema (~14 Mya in the great ape lineage): Great apes demonstrate mirror self-recognition (Gallup, 1970), indicating a self-schema that integrates body schema, spatial schema, and attention schema into a unified model of the self as a persistent entity distinct from others. The great ape lineage (Hominidae) diverged from other primates approximately 14–16 Mya; orangutans, chimpanzees, bonobos, and gorillas pass the mirror test while most monkeys do not. Mirror self-recognition has also been reported in Asian elephants (Plotnik et al., 2006), Eurasian magpies (Prior et al., 2008), and possibly cleaner wrasse (Kohda et al., 2019), indicating that self-recognition capacities have evolved convergently across distantly related lineages. This complicates the use of MSR as a strict marker for the ~14 Mya dating; the RSP model predicts that these species should show correspondingly limited recursive social prediction depth, which current evidence supports. The self-schema is the architectural bridge between Level 2 and Level 3: without it, the recursive social-predictive loop has no "self" to model.

This sequential emergence makes a testable prediction: in infant development, the schemas should appear in roughly the same order as their evolutionary emergence. Body schema first (present at birth), spatial schema next (reaching and orienting at 3–4 months), affective-homeostatic regulation (6–9 months), attention schema (joint attention at 9–12 months), and self-schema (mirror self-recognition at 18–24 months). The developmental evidence confirms this sequence (Amsterdam, 1972; Meltzoff, 2007; Rochat, 2003).

Level 3: The Hominin Revolution

Level 3—recursive social prediction—emerged uniquely in the hominin lineage, driven by escalating social complexity. Dunbar's (1998) Social Brain Hypothesis provides the quantitative framework: across primates, neocortical volume correlates not with ecological variables (diet, territory size) but with social group size. The demands of tracking alliances, rivalries, debts, reputations, and deceptions across increasingly large groups selected for neural architectures capable of recursive modeling: *If I share with her, he will see, and he might retaliate because he thinks I'm favoring her over him.*

The archaeological record allows us to trace this development with surprising specificity:

7–3 Mya (early hominins, groups of 20–50): *Sahelanthropus*, *Ardipithecus*, and early *Australopithecines* lived in small groups in mixed woodland-savanna environments. Cranial capacities of 350–450cc supported Level 2 schemas and basic social cognition: recognizing individuals, remembering alliances, tracking simple hierarchies. Level 3 was nascent at best.

3.3–2 Mya (Lomekwian and Oldowan tools): The earliest known stone tools (Lomekwi 3, Kenya, ~3.3 Mya) and the more widespread Oldowan industry (~2.6 Mya) of *Homo habilis* (brain ~600–750cc) required planned flaking sequences and, critically, *teaching* — or at minimum, sustained social observation. Active teaching requires modeling the learner's state of knowledge: "They don't yet know how to strike at this angle." This is second-order social prediction—Level 3a (conceptual categorization applied to other minds) emerging under selection pressure.

1.8 Mya–800 Kya (Homo erectus, groups of 80–100): *Homo erectus* represents a cognitive leap: brain sizes of 800–1100cc, Acheulean hand axes requiring extended multi-step production sequences (*chaînes opératoires*), evidence of controlled fire use (Wonderwerk Cave, South Africa), and migration out of Africa. Tracking alliances, debts, and deceptions across groups of 80–100 individuals demands recursive modeling. The strange loop was beginning to operate in earnest.

800–300 Kya (archaic humans, groups of 100–130): *Homo heidelbergensis* and early *Homo sapiens* (brain 1100–1400cc) developed cooperative big-game hunting requiring real-time social coordination. The Schöningen spears (Germany, ~300 Kya)—carefully crafted throwing spears—imply planned group hunts in which each participant must predict what their hunting partners will do, predict that their partners predict what they will do, and coordinate accordingly. This is full recursive social prediction operating under lethal pressure, where failure of the strange loop means death.

300–70 Kya (Homo sapiens, behavioral modernity): Anatomically modern humans emerged in Africa with brain sizes of 1200–1500cc and, critically, reorganized prefrontal and temporal cortices. Symbolic behavior appeared: ochre pigments at Blombos Cave (~100 Kya), shell beads at Skhul Cave (~100 Kya), geometric engravings. Symbols require shared meaning—the ability to represent something *to someone*, knowing they will understand it as you intend. This is Level 3a (conceptual categorization) reaching its full potential: meaning is now publicly shared and recursively interpreted.

The behavioral revolution of the Upper Paleolithic (~70–40 Kya), with its blade technologies, composite tools, personal ornaments, long-distance trade networks, and the earliest cave art (Chauvet, ~36 Kya; Lascaux, ~17 Kya), required *cumulative culture*: knowledge transmitted across generations through teaching, imitation, and symbolic communication. Cumulative culture demands full Level 3: you must model what others know, predict what they need to learn, and represent your own knowledge in transmissible form.

Beyond Dunbar's Number: When Biology Required Culture

Perhaps the most striking evidence for the RSP model comes from what happened when human societies exceeded the ~150-person limit of direct recursive social prediction (Dunbar's number). Rather than evolving larger brains, humans invented *cultural technologies* that extended Level 3 beyond its biological substrate:

Role-based prediction: In agricultural villages (beginning ~12 Kya in the Fertile Crescent), individuals could no longer maintain recursive models of every community member. The solution was categorical social prediction: instead of modeling "Ugg the specific person," you model "the chief," "the healer," "the stranger," social roles with predictable behaviors.

Externalized social memory: Writing (Sumer, ~3200 BCE) externalized social prediction into durable media. Law codes (Ur-Nammu, ~2100 BCE; Hammurabi, ~1750 BCE) codified social predictions as explicit rules: "If you do X, consequence Y will follow." Markets and currency created abstract social prediction: modeling the behavior of strangers through standardized exchange.

Institutional prediction: Cities and civilizations developed increasingly sophisticated cultural institutions (legal systems, religions, bureaucracies, democratic assemblies), each of which functions as a social prediction technology, allowing individuals to coordinate behavior with thousands or millions of people they will never meet.

This sequence reveals something fundamental about the RSP architecture: it has *limits*. The biological substrate of Level 3 maxes out at roughly 150 recursive social models. When social complexity exceeded this limit, humans did not evolve new neural architecture. They built cultural scaffolding for the existing architecture. The fact that every complex society has independently invented institutions, codified laws, and symbolic systems for extending social prediction beyond face-to-face interaction strongly suggests that the RSP architecture is real, that it has discoverable constraints, and that it shapes the structure of human civilization itself.

Convergent Evolution and Alternative Selection Pressures

The mammalian-centric narrative above requires an important caveat. Corvids (crows, jays, ravens) show cognitive sophistication rivaling great apes: tool manufacture in New Caledonian crows, future planning in Western scrub-jays, episodic-like memory, and possible theory of mind (Emery & Clayton, 2004). Yet they live in relatively small social groups. This suggests that ecological demands (extractive foraging, tool use) can independently drive cognitive sophistication, as the ecological intelligence hypothesis proposes (DeCasien et al., 2017, found that frugivory is a stronger predictor of primate brain size than social group size once phylogeny is controlled).

Cephalopods present an even sharper test case. Octopuses possess approximately 500 million neurons, comparable to a dog, organized in a radically different architecture: two-thirds of their neurons reside in the arms rather than a centralized brain (Mather, 2008; Fiorito & Scotto, 1992). They demonstrate sophisticated problem-solving (opening jars, navigating mazes, using coconut shells as tools), rapid chromatophore-driven camouflage requiring real-time body-environment modeling, and observational learning. Yet octopuses are largely solitary, with minimal social interaction beyond mating. Under the RSP model, this predicts high Level 1 capacity (valenced states driving approach/avoidance) and rich Level 2 schemas (an exceptionally complex body schema for coordinating eight semi-autonomous arms, spatial schema for den navigation and hunting) without Level 3 recursive social prediction. The available evidence supports exactly this: octopuses show behavioral flexibility and environmental intelligence rivaling social mammals, yet no evidence of mirror self-recognition, theory of mind, or recursive self-modeling. They are, in RSP terms, a system with high $\mathcal{I}_1 + \mathcal{I}_2$ (Levels 1–2 intelligence) but negligible $\mathcal{I}_{\text{rec}}$ (recursive depth) — a prediction formalized in Appendix H (§H.5, Prediction H.4).

The RSP model's strongest claim is therefore not that *all* cognitive evolution tracks social complexity, but that the specifically *recursive, self-referential* character of human consciousness—the strange loop—was

shaped by social demands. Corvids may have evolved sophisticated Level 2 schemas and proto-Level 3 capacities through ecological rather than social pressures, which is consistent with the RSP architecture (ecological problems also require predictive modeling) but challenges the exclusively social selection narrative. Similarly, cooperative hunting in wolves, orcas, and lions achieves behavioral coordination through simpler mechanisms (role-based heuristics, shared attention to prey) without necessarily requiring full recursive mental state attribution.

The Evolutionary Argument Summarized

The RSP model predicts that consciousness should have evolved in precisely those lineages where social complexity demanded recursive self-modeling, should have elaborated in proportion to social group size, and should have produced cultural extensions when biological capacity reached its limits. The paleoanthropological record confirms all three predictions:

1. **Level 1** (core affect) is conserved across all vertebrates, reflecting its ~450-million-year heritage.
2. **Level 2** (core schemas) emerged sequentially in the mammalian lineage, with each schema corresponding to a specific ecological and social demand.
3. **Level 3** (recursive social prediction) elaborated specifically in the hominin lineage, correlating with social group size (Dunbar, 1998), not with brain size per se.
4. **Cultural extensions** of Level 3 (language, institutions, writing, law) emerged precisely when social groups exceeded the biological capacity for direct recursive modeling.

No rival theory of consciousness generates this pattern of predictions. Integrated Information Theory (IIT) would predict consciousness proportional to neural complexity (Φ) regardless of social demands. Global Workspace Theory would predict consciousness proportional to information broadcasting capacity regardless of social structure. Only the RSP model predicts that the evolution of consciousness should track *social complexity specifically*, a prediction the comparative and archaeological evidence consistently supports.

The Computational Substrate: What Evolution Built

The human brain comprises approximately eighty-six billion neurons, each connecting to thousands of others through synapses to create between one hundred trillion and five hundred trillion total synaptic connections. Parallel processing across these networks produces computational throughput on the order of 10^{16} operations per second. The essential point for the RSP model is that none of this computational activity is itself conscious. Consciousness is not an observation of neural processing; it is what it is like, from the inside, to *be* a brain whose computations achieve the specific integrative pattern described by the three-level architecture.

Why This Matters for the RSP Model

This substrate-level reality supports the meaningful integration framework directly. By the time any content reaches consciousness, it has already been extensively processed, categorized, and integrated across all three levels: there is no stage at which "raw, pre-conceptual phenomenal qualities" exist to be

subsequently interpreted. The explanatory gap appears unbridgeable only when this intermediate processing is overlooked; once recognized, the question shifts from how physical events transform into subjective qualities to what organizational patterns of computation constitute experience from the first-person perspective. The RSP architecture identifies that organizational pattern: what reaches awareness is specifically what has been integrated through core affect (Level 1), organized through schemas (Level 2), categorized through concepts, and woven into the recursive self-model (Level 3). The brain's computational power is necessary but not sufficient; what matters is how that computation is organized through the three-level architecture.

The Information-Theoretic Parallel: From Data to Self-Referential Computation

The three-level RSP architecture maps onto a transition well understood in information theory: from raw data, through structured information, to self-referential computation. Level 1 core affect functions as continuous, undifferentiated signal, analogous to unprocessed sensor readings. Level 2 schemas impose categorical structure on that signal, transforming it into relationally organized representations. Level 3 introduces self-reference: the prediction-error loop (predict future self-state, compare to actual, update model) is structurally identical to predictive coding, and the recursive quality — a system modeling itself modeling itself — is what emerges when computation takes itself as input.

This parallel matters philosophically because the transition from data to structured information to computation generates no conceptual mystery in information theory. The puzzle arises only at the self-referential step. If consciousness follows this same architectural pattern, then the hard problem is not about an ineffable extra ingredient but about what happens when a sufficiently complex information-processing system becomes recursively self-modeling. The explanatory gap falls not between physics and phenomenology but between computation and *self-referential* computation — a gap that is substantive but not metaphysical.

One crucial disanalogy strengthens the argument. Unlike genuinely neutral binary data, Level 1 core affect is already minimally meaningful: already valenced, already carrying survival relevance. Consciousness therefore emerges not from meaningless substrate but from something *already minimally meaningful* being progressively structured (Level 2) and then recursively modeled (Level 3). This is why the RSP model avoids both the hard problem (how does consciousness emerge from something utterly non-conscious?) and the combination problem facing panpsychism (how do micro-experiences combine into macro-experience?). Core affect is the biological ground floor — already valenced, already meaningful in the minimal sense that matters for survival — upon which schemas and recursive modeling build the full architecture of phenomenal experience.

The Developmental Sequence: Consciousness Emerging Socially

This social-predictive account explains developmental patterns:

0-18 months: Infants have core affect organized through developing schemas (body, spatial) but lack sophisticated self-models. They have experience but not yet self-conscious experience. They cannot yet model themselves as others see them.

18-24 months: Mirror self-recognition emerges. Infants recognize themselves in mirrors, indicating a developing self-model. They begin to understand themselves as entities that can be perceived by others.

4-5 years: Explicit theory of mind develops. Children pass false-belief tasks, demonstrating they understand that others can have different knowledge and beliefs (Wellman et al., 2001; implicit precursors appear earlier, around 15 months). Crucially, this requires both modeling others as having mental states AND modeling oneself as having mental states that differ from others'. The recursive loop begins operating: "I know X, but Mom doesn't know X, so if I want her to know X, I must tell her."

4-7 years: Autobiographical memory develops. Children construct narratives about their past selves, creating continuity. The self-model extends across time. They can predict their future mental states and remember past mental states as belonging to the same "I."

7+ years: Sophistication increases. Children become better at taking others' perspectives, understanding complex social dynamics, predicting social consequences. The recursive loops deepen: "They think that I think that they think..."

As these capacities develop, consciousness becomes richer. The child develops a stronger sense of self as continuous entity, becomes able to think about their own thinking, to reflect on experience, to imagine how others see them. The strange loop elaborates and the self-model integrates with increasingly sophisticated social predictions.

Why this ordering is mathematically necessary. The formal appendixes show that the Level 1 $\rightarrow$ 2 $\rightarrow$ 3 developmental sequence is not merely an empirical observation but an architectural constraint of the RSP framework. The free energy formulation (Appendix B) requires F_1 to be functional before F_2 can minimize, because Level 2 prediction errors $\varepsilon_2 = \mu_1 - g_2(\mu_3)$ depend on stable Level 1 expectations μ_1 . The reinforcement learning formulation (Appendix C) requires a reward signal (valence, Level 1) before state representations (schemas, Level 2) can be learned through unsupervised exploration — play is the infant's exploration policy, driven by intrinsic reward. The control-theoretic formulation (Appendix D) requires inner control loops to be stable before outer loops can be commissioned — the cascade design principle $\tau_1 < \tau_2 < \tau_3$ implies a construction order matching the developmental sequence. In all three formalisms, the architecture cannot be assembled top-down or simultaneously; it must be built from the bottom up, exactly as observed in human development.

Why Animals Have Simpler Consciousness

The evolutionary arc described above explains the gradations of consciousness we observe across phylogeny. Each species' consciousness reflects the level of the RSP architecture its lineage has reached, and critically, the social complexity that drove its development:

Simple vertebrates (fish, amphibians, reptiles): These animals possess Level 1 core affect—the phylogenetically ancient subcortical systems generating valence and arousal (~450 Mya heritage)—and rudimentary Level 2 schemas (body, spatial). Their experience is valenced and organized but not recursive. A fish navigating a reef has affective responses to threats and rewards organized through body and spatial schemas, but does not model itself as an entity with a persistent identity.

Non-social mammals (cats, bears, hedgehogs): These species have richer Level 2 schemas including proprioceptive body schema, spatial schema (hippocampal cognitive maps), and affective-homeostatic schema (the mammalian innovations of ~250–150 Mya). They can predict their own immediate states. A cat predicts its own hunger, fatigue, and interest in prey, enabling phenomenology richer than mere affective response. Even solitary mammals retain the neural architecture for self-modeling that originally evolved for social purposes, since their ancestors were social.

Social mammals (elephants, dolphins, wolves, some primates): These animals have sophisticated attention schemas and rudimentary self-schemas, enabling recognition of self in mirrors (in some species), longer-term memory integration, and basic social prediction. Elephant herds, dolphin pods, and wolf packs require modeling other individuals' likely behavior—a proto-Level 3 capacity. But the recursive depth is limited: they predict what others will do, yet show limited evidence of predicting what others *think about them*.

Great apes (chimpanzees, bonobos, orangutans, gorillas): Great apes possess full self-schemas (mirror self-recognition, emerging in the great ape lineage ~14 Mya) and theory-of-mind precursors. They can model themselves from others' perspectives to some degree and engage in tactical deception (Byrne & Whiten, 1988). Their social groups of 30–60 individuals demand recursive social prediction, but their Level 3 remains limited by the absence of linguistic concepts and cumulative culture. A chimpanzee can model "he thinks I'm submissive," but cannot model "he thinks I think he thinks I'm submissive."

Humans: Full recursive self-modeling with linguistic elaboration, operating within social groups that have exceeded Dunbar's number through cultural extensions. We can model ourselves modeling others modeling us, ad infinitum. We can construct elaborate narratives about our past and future selves. We can use culturally learned concepts to create fine-grained experiential distinctions unavailable to any other species. The richness of human consciousness is not a product of brain size alone. It is the product of the recursive social prediction architecture calibrated through years of social interaction, elaborated through language, and extended through cultural institutions.

The richness of consciousness thus correlates with: (1) sophistication of core schemas, (2) depth of recursive self-modeling, (3) complexity of social environment and demands, and (4) availability of cultural concepts for elaboration. This is not merely a descriptive taxonomy. It is a set of predictions that the comparative evidence confirms.

Natural Experiments: Social Deprivation and Consciousness Development

The strongest empirical support for Level 3's social-predictive architecture comes from tragic natural experiments: cases where human biology is intact but social interaction is absent or severely curtailed during development. If the RSP model is correct, these individuals should show preserved Levels 1-2 (core affect and basic schemas) with selectively impoverished Level 3 (recursive self-modeling, theory of mind, emotional granularity). This is precisely what we observe.

Genie Wiley (Curtiss, 1977): Confined to a small bedroom from approximately 20 months of age until age 13 with near-total social isolation — strapped to a potty chair during the day and restrained in a sleeping

bag at night — Genie displayed clear core affect. She showed fear, pleasure, anger, and attachment upon rescue (Level 1 intact). She had functional body and spatial schemas; she could walk, navigate rooms, manipulate objects (Level 2 intact). But her Level 3 capacities were devastatingly impaired: she never developed normal theory of mind, her emotional granularity remained severely limited (she could express basic affects but struggled to differentiate complex social emotions), her autobiographical narrative was fragmentary, and she showed minimal capacity for recursive self-modeling. She could not readily imagine herself from others' perspectives or engage in the kind of social prediction that characterizes normal adult consciousness. Despite years of intensive rehabilitation, these Level 3 capacities never fully developed, consistent with a critical period for establishing the recursive social-predictive loop.

Romanian orphanage studies (Nelson, Fox & Zeanah, 2014): The Bucharest Early Intervention Project studied children raised in Romanian institutions with minimal caregiver interaction, fed and sheltered but rarely held, spoken to, or engaged socially. These children showed a striking dissociation between levels. Core affect was present: they displayed distress, pleasure, and arousal responses (Level 1). Basic schemas functioned: they developed body awareness and spatial navigation (Level 2). But Level 3 was profoundly impaired: they showed reduced cortical gray matter in regions associated with self-modeling and social cognition, delayed or absent mirror self-recognition, poor theory of mind, and a characteristic pattern of "indiscriminate friendliness"—approaching strangers without genuine social modeling, as though the social-prediction circuitry was reaching for input it had never been calibrated on. Crucially, children randomized to placement in foster care before age 2 recovered significantly more Level 3 function than those placed later, suggesting a sensitive period during which the recursive loop must receive social input to properly develop.

Feral children (Itard, 1801; Singh & Zingg, 1942; reports on Oxana Malaya): Children raised by animals or in extreme isolation consistently show the same dissociation the RSP model predicts. Victor of Aveyron, studied by physician Jean-Marc Itard in the early 19th century, displayed intact approach/avoidance behaviors and basic sensory-motor schemas but showed no evidence of self-conscious reflection, emotional differentiation beyond basic affects, or recursive social cognition. Oxana Malaya, raised among dogs in Ukraine until age 8, had functional body schemas adapted to quadrupedal movement but lacked self-consciousness, emotional vocabulary, and theory of mind upon rescue. In each case: Levels 1-2 present, Level 3 absent or minimal, exactly what the architecture predicts for organisms without social-predictive input.

Harlow's isolation experiments (Harlow, 1958; Harlow & Harlow, 1965): While ethically troubling by modern standards, Harry Harlow's studies of rhesus monkeys raised in total social isolation provide controlled evidence for the same pattern. Isolated monkeys developed intact sensory processing and motor schemas (Level 2) but showed devastating deficits in social cognition, emotional self-regulation, and behavioral flexibility (Level 3). They could not read social cues, could not predict others' behavior, and displayed stereotyped rather than flexible responses, consistent with a self-model that was never calibrated through social interaction. Monkeys isolated for shorter periods showed partial recovery, again suggesting a sensitive period for Level 3 development.

What this pattern reveals: Across species, across centuries, across wildly different circumstances, the same broad dissociation appears. Remove social interaction while leaving biology intact, and the most prominent deficits fall on recursive social prediction (Level 3) while core affect (Level 1) and basic schemas (Level 2) remain relatively preserved. The dissociation is not perfectly clean. Genie had motor difficulties, Harlow's monkeys showed stereotyped behaviors suggesting some Level 2 disruption, and Romanian orphans showed broad cognitive delays beyond social cognition. But the *pattern* is clear: the Level 3 deficits are disproportionately severe relative to Levels 1–2, not uniformly distributed across all levels. This is not what competing theories predict as cleanly. Integrated Information Theory would predict consciousness deficits proportional to reduced neural complexity, but these children have structurally normal brains at birth. Global Workspace Theory would predict deficits proportional to reduced information broadcasting, but the deficit is specifically social-recursive, not broadly cognitive. The RSP model uniquely predicts the selective Level 3 impairment that natural experiments consistently reveal.

The recovery data is equally telling: earlier social intervention produces better Level 3 outcomes. This is consistent with a critical period during which the recursive loop must receive social calibration. Others' responses to the child's behavior become the training signal for the self-model. Without this signal, the loop cannot bootstrap itself. The architecture is present but uninitiated, like a language faculty that was never exposed to language.

Disorders of Consciousness: The Most Controlled Natural Experiments

The social deprivation cases above involve confounding variables: malnutrition, trauma, sensory deprivation alongside social deprivation. Disorders of consciousness offer something closer to controlled experiments, focal disruptions of specific neural substrates that can be mapped with neuroimaging to particular levels of the RSP architecture.

Vegetative state (VS). Patients in the vegetative state retain brainstem reflexes, sleep-wake cycles, and autonomic regulation, the neural infrastructure of Level 1 core affect, while showing no purposeful behavior or communication (Giacino et al., 2014). Under the RSP model, VS represents preserved Level 1 with disrupted Level 2–3 integration. The architecture predicts that VS patients should exhibit interoceptive processing, including autonomic responses to painful stimuli, without schema-organized experience. This is precisely what clinical observation confirms: VS patients show heart rate and blood pressure changes in response to noxious stimulation, but these responses lack the organized, contextually modulated character that schema-level processing would produce.

Minimally conscious state (MCS). Patients in MCS show intermittent but reproducible evidence of awareness: visual tracking, localization of sound, command-following that appears and disappears across assessments (Giacino et al., 2014). Under the RSP model, MCS represents partially functioning Level 2 schemas with inconsistent Level 3 engagement. The prediction follows directly: MCS patients should show variable but detectable schema-level processing, such as tracking objects (spatial schema) and localizing sounds (attention schema), without consistent recursive self-modeling. The intermittent quality of MCS awareness is itself informative. It suggests that the Level 2 schemas are structurally present but lack the

stable integration needed for sustained conscious experience, a pattern consistent with fluctuating thalamocortical connectivity rather than permanent structural loss.

Locked-in syndrome. Patients with locked-in syndrome retain full consciousness despite near-total motor paralysis, typically resulting from ventral pontine lesions that spare the reticular activating system and cortex. Under the RSP model, all three levels remain intact; the deficit lies entirely in motor output, not in the RSP architecture. This is a critical control case: consciousness without behavioral expression, confirming that the RSP architecture is internal and computational, not defined by behavioral capacity.

Covert awareness in vegetative state. Owen et al. (2006) demonstrated that a patient clinically diagnosed as vegetative — not locked-in — could perform mental imagery tasks on command, detected via fMRI, revealing intact higher-level processing behind a behaviorally unresponsive exterior. This landmark finding exposed a population of misdiagnosed VS patients who retain covert awareness, and it underscores a critical distinction for the RSP model: behavioral unresponsiveness does not entail absence of the RSP architecture. Under RSP, these patients likely retain Level 2 schema processing (the mental imagery tasks activated spatial and body schemas) and possibly Level 3 recursive capacity, despite meeting clinical criteria for vegetative state.

TMS-EEG complexity measures. Casali et al. (2013) demonstrated that the Perturbational Complexity Index (PCI), which measures the spatiotemporal complexity of cortical responses to transcranial magnetic stimulation, reliably discriminates VS, MCS, and fully conscious states with high sensitivity and specificity. Under the RSP model, PCI tracks the degree of cross-level integration: low PCI reflects Level 1 processing only, producing simple, stereotyped cortical responses; high PCI reflects multi-level integration with the recursive complexity characteristic of Levels 2 and 3. The prediction: PCI should correlate with the number of active RSP levels, not merely with "global workspace ignition." This interpretation remains to be directly tested against RSP-specific level predictions, as PCI values cluster into ranges that map onto the clinical distinctions the RSP model would predict.

Anesthesia. General anesthesia offers a pharmacologically controlled window into the RSP architecture. Different anesthetic agents disrupt different neural substrates in ways that map onto the level structure. Propofol primarily disrupts thalamocortical connectivity, the substrate for Level 2 schema integration, producing dreamless unconsciousness in which all three levels go offline in a characteristic sequence (this sequential shutdown is an RSP prediction, not yet empirically established as a level-specific sequence). Ketamine, by contrast, disrupts cortical-cortical connectivity while partially preserving thalamocortical loops, corresponding to disrupted Level 3 recursive processing with partially preserved Level 1-2 function. The phenomenological consequences align with this architectural analysis: propofol produces an absence of experience, while ketamine produces dissociative states, distorted but not absent experience in which core affect and fragmented schemas persist without coherent recursive self-modeling. The RSP model predicts that different anesthetic agents should produce distinct patterns of level-specific disruption with correspondingly distinct phenomenological signatures, a prediction that pharmacological studies of subjective experience under anesthesia are beginning to confirm.

One complication: Sarasso et al. (2015) showed that ketamine produces high PCI values comparable to wakefulness, despite producing dissociative unconsciousness. Under RSP, this is explained by PCI measuring cortical complexity (which ketamine preserves) rather than recursive self-modeling specifically (which ketamine disrupts). PCI tracks the richness of Level 2 schema processing, not the recursive depth of Level 3b.

Connecting existing findings to RSP levels. The Owen et al. (2006) finding discussed above is particularly informative: the specific mental imagery tasks the patient performed — imagining playing tennis (activating supplementary motor area) and navigating her house (activating parahippocampal gyrus) — provide direct evidence for preserved Level 2 schema processing (body schema and spatial schema, respectively) in the absence of behavioral indicators. Casali et al.'s (2013) PCI values discriminate consciousness states with high accuracy: low PCI in vegetative state (consistent with Level 1-only processing), intermediate PCI in minimally conscious state (consistent with partial Level 2), and high PCI in conscious states (consistent with multi-level integration). These findings provide the most controlled existing evidence for RSP's level-specific architecture.

Why disorders of consciousness matter for RSP. These clinical populations provide the most direct tests of the RSP model's level-specific predictions. Social deprivation cases demonstrate that Level 3 requires social input to develop; disorders of consciousness demonstrate that the levels can be independently disrupted in the mature brain. The convergence of evidence from both sources, developmental and neurological, targeting the same architectural distinctions from opposite directions, strengthens the case that the three-level structure is not merely a theoretical convenience but reflects genuine joints in the organization of conscious experience.

VI. Objections and Replies: Addressing the Social-Predictive Account

The evolutionary, developmental, and clinical evidence assembled in the previous section supports the RSP architecture empirically. But does the framework hold up under philosophical scrutiny? I now turn to four traditional puzzles and show that each dissolves under the meaningful integration account.

The Privacy of Qualia Dissolved

Here is the traditional problem in its sharpest form: if qualia are raw, private, intrinsic properties of individual experiences, how could we ever know whether your red matches my red? We might use “red” to refer to completely different experiences, and communication about color would rest on nothing solid.

This is sometimes called the “inverted spectrum” problem: perhaps your color experiences are systematically inverted relative to mine (what you experience as red, I experience as green), but we've both learned to call the long-wavelength color “red,” so we never detect the inversion.

Under meaningful integration, this problem dissolves. There is no “raw red experience” to compare. What exists instead:

Biological similarity: Our visual systems share structural features, so red surfaces produce similar activation patterns in both of us. Similar core affects follow—arousal patterns, attentional salience.

Cultural concepts: We learn the concept RED from shared language and cultural practices. It carries semantic associations (blood, roses, stop signs), emotional valences (danger, passion), and functional roles (warning signals) that speakers of a language largely hold in common.

Personal history: Your specific encounters with red things create your unique associative network. Your red includes memories of your grandmother’s red coat, your feelings about stop signs, your experience of embarrassment (going red).

Integrated phenomenology: When you see something red, your experience weaves together biological response, cultural concept, and personal associations. So does mine. Our experiences are neither identical (different histories) nor entirely private (shared biology and culture). They are **intersubjectively similar**—similar enough to ground communication, different enough to remain individually unique.

We communicate about red successfully not despite but *because* the experience is meaningfully integrated. Meaning provides the overlap. We share enough biological response and conceptual knowledge to coordinate our use of “red,” even as our precise phenomenologies differ.

The inverted spectrum scenario becomes incoherent on this view. You cannot have systematically inverted color experiences while learning color concepts normally, because concept learning depends on pattern-matching across multiple modalities. “Red” means dangerous/hot/stop/blood/roses—a semantic network that constrains phenomenology. If your long-wavelength experience matched my green-associations, you would fail to learn the concept correctly.

The Problem of Other Minds Dissolved

How can I know that other beings have conscious experiences? Perhaps every other human is a “philosophical zombie”—physically identical to a conscious being but lacking any inner experience. Perhaps animals lack consciousness entirely.

This problem assumes consciousness is an all-or-nothing property, a mysterious inner light that is either present or absent, which only I can verify in my own case. Under meaningful integration, consciousness is not binary but a matter of degree and kind.

The relevant questions become:

- What is the organism’s capacity for core affect? (Do they have valence/arousal systems?).
- What conceptual categories do they possess? (What differentiations can they make?).
- How sophisticated are their predictive models? (How far ahead do they simulate?).
- What is the scope of their integration? (What modalities and timescales are integrated?).

These are empirical questions, answerable through behavior, neuroscience, and comparative psychology. We do not need telepathic access to another’s raw qualia. We need to understand their functional architecture.

For other humans: We share biology, culture, and language, giving us strong reason to believe their meaningful integrations resemble ours. They learn the same emotion concepts, respond to similar situations with similar patterns, and report experiences that make sense given their situation. Their consciousness is not metaphysically mysterious. It is expectable given our shared architecture.

For animals: The question is not “Do bats have raw bat-qualia?” but “What meaningful integrations do bats perform?” Bats clearly have core affect (approach/avoidance, pleasure/distress). They likely have some conceptual categories (prey/non-prey, roost-mate/stranger). Their predictive models incorporate echolocation in ways ours don’t. Their consciousness differs from ours not by possessing different raw qualia but by performing different integrations suited to different ecological niches.

For AI systems: Current artificial systems lack the biological substrate that generates core affect. They may possess conceptual categories and predictive models, but without valenced arousal states grounded in homeostatic needs, they lack a crucial component of meaningful integration. This does not rule out machine consciousness in principle—it clarifies what would be required: synthetic affects, self-modeling, and conceptual integration.

The Explanatory Gap Dissolved

There seems to be an unbridgeable “explanatory gap” (Levine, 1983) between physical processes (neural firing patterns) and phenomenal experience (what it is like to see red). We can exhaustively describe neural activity without thereby explaining why that activity is accompanied by experience.

The gap looks unbridgeable if we assume experiences are raw qualia—intrinsic properties that float free of functional role. No amount of functional description explains why a particular functional state should have a particular raw feel.

But if experiences are meaningful integrations, the gap dissolves. The question shifts: the task is no longer to explain how neurons produce raw, meaning-independent feels, but to explain how biological processes implement meaningful integration.

This remains challenging—but it is the kind of challenge science routinely addresses. How do lower-level processes implement higher-level properties? How does water arise from H₂O molecules? How does life arise from chemistry? How does meaning arise from computation?

Under the RSP framework, asking “why should meaningful integration feel like anything at all?” presupposes that phenomenology could be absent from the recursive process that constitutes it—a presupposition the framework rejects as a category error. The question is dissolved, not merely narrowed. Several traditional pseudoproblems fall away with it:

- Not: “Why does red feel like *this*?” (There is no meaning-independent “this”).
- Not: “How can we compare your red to mine?” (Experiences are intersubjectively structured).
- Not: “Could zombies exist?” (Consciousness is functional integration, not ghostly extra).

The question “why is meaningful integration accompanied by felt experience?” mistakes a constitutive relationship for a causal one—like asking “why does rotation rotate?” The explanatory gap depended on

treating phenomenology as ontologically separate from functional architecture. Once that separation is rejected, the gap closes.

Cultural Variation as Prediction, Not Puzzle

The traditional framework makes cultural variation in consciousness puzzling. If emotions are biological universals with the same raw phenomenology everywhere, why do cultures categorize them so differently?

Standard responses:

- Same emotions, different labels (but this fails empirically—concepts don’t translate).
- Same core emotions plus cultural overlay (but which emotions are core?).
- Cultural display rules overlay universal feelings (but this can’t explain variation in reported experience).

Under meaningful integration, cultural variation is a central prediction, not a puzzle. Different cultures furnish different conceptual categories, which constitute genuinely different feelings.

Consider Japanese *amae*—roughly, pleasurable dependence on another’s benevolence. English speakers do not have quite this feeling because English does not have quite this concept. The biological core affect might be similar (pleasant low-arousal with affiliative motivation), but the specific feeling emerges from categorizing this core affect using the concept *amae* with its particular meanings in Japanese social life.

This is not “just semantic.” The concept shapes phenomenology:

- What situations trigger the feeling (expecting benevolence).
- What action tendencies it includes (seeking caregiving).
- What moral valence it carries (positive dependence vs. problematic neediness).
- What developmental arc it follows (appropriately expressed in childhood, maturely expressed in intimate relationships).

English speakers experiencing similar core affect might categorize it as “feeling needy” (negative), “being cared for” (neutral), or “feeling close” (positive)—each creating different phenomenology because different conceptual framings integrate the core affect differently.

This accounts for otherwise puzzling findings:

- Emotions don’t translate cleanly across languages.
- Emotion recognition is better within-culture than across-culture.
- Emotional granularity varies across individuals and develops over the lifespan.
- Emotion concepts shape neural activity patterns, not just vice versa.

Rather than assuming universal raw feelings dressed in cultural clothing, I argue that concepts partially constitute feelings. Culture does not overlay consciousness—it partially constructs it.

VII. What It's Like to Be a Bat: The Question Reframed

Why Nagel's Question Misfires

Section IV established why “What is it like to be a bat?” misfires as a question: it presupposes raw, meaning-independent phenomenology that could be accessed if only we had the right vantage point. The bat question dissolves under the RSP framework because there is no substrate-neutral “what it is like” separable from the bat's specific architecture. But the dissolution is not merely negative. We can now ask productive questions about bat consciousness by examining the schemas through which a bat's experience is organized.

What To Ask Instead

Instead of “What is it like to be a bat?” I propose asking:

Architectural questions:

- What core affects do bats experience? (Evidence: approach/avoidance behaviors, distress calls, pleasure responses).
- What conceptual categories do bats possess? (Evidence: discrimination learning, behavioral flexibility).
- How sophisticated are bat predictive models? (Evidence: navigation planning, prey anticipation).
- What is the scope of temporal integration? (Evidence: episodic-like memory, delayed gratification).
- What self-other models do bats construct? (Evidence: recognition, social learning, cooperation).

Ecological questions:

- How does echolocation shape bat predictive models differently from vision?
- What affordances does echolocation provide that vision doesn't, and vice versa?
- How do bat social structures shape what categories bats learn?
- What evolutionary pressures shaped bat consciousness in directions different from primates?

Comparative questions:

- How does bat consciousness compare to dolphin consciousness (also echolocating)?
- How does it compare to bird consciousness (also flying, also navigating 3D space)?
- What aspects of consciousness are convergently evolved across mammals?
- What aspects are lineage-specific or niche-specific?

What We Can and Cannot Know

Under this reframing, some questions become answerable while others dissolve:

We can know:

- Whether bats have core affect (yes—behavioral evidence is strong: approach/avoidance, distress vocalizations, pleasure responses)

- What core schemas bats possess:
 - **Body schema:** Highly sophisticated for flight, three-dimensional navigation.
 - **Spatial schema:** Allocentric cognitive maps for roost locations, foraging routes.
 - **Attention schema:** Models where conspecifics are directing attention (essential for social roosting).
 - **Self-schema:** Likely present but simpler than primate self-models.
 - **Affective-homeostatic schema:** Monitors hunger, fatigue, threat.
- How bat schemas differ structurally from human schemas:
 - Body schema integrates echolocation (temporal/auditory) rather than vision (spatial/visual).
 - Spatial schema emphasizes 3D volumetric space over 2D surface navigation.
 - Attention schema tracks conspecifics but likely lacks theory-of-mind sophistication.
- That bat consciousness is genuine structured experience organized through these schemas.

We cannot know (and don't need to):

- What raw bat-echolocation-qualia “feel like” (no such thing exists).
- Whether bat-red matches human-red (concepts and schemas don't map this way).
- Whether bats have “true” consciousness or mere “as-if” consciousness (false dichotomy based on raw qualia assumption).

The key insight is this: bat consciousness differs from human consciousness not in having different raw qualia but in having differently structured schemas—body schema organized around echolocation, spatial schema optimized for 3D flight—combined with simpler conceptual elaboration (fewer learned categories, no linguistic concepts).

We can investigate bat consciousness empirically by studying:

- What behavioral discriminations bats make (revealing their conceptual categories).
- How bat neural networks organize sensory information (revealing schematic structure).
- What predictions bats generate (revealing forward models).
- How bats respond to novelty, threat, reward (revealing affective organization).

The question dissolves:

- “What is it like to be a bat?” assumes a determinate fact about raw bat-phenomenology accessible in principle.
- But phenomenology is constituted by meaningful integration through schemas and concepts.
- Bat meaningful integrations differ from human integrations at both schematic and conceptual levels.
- There's no neutral standpoint from which to compare raw feels because raw feels don't exist independently of schematic organization.
- The right question is about functional architecture and schema structure, which we can investigate.

A Concrete Example: Echolocation Through Schemas

Consider how a bat “experiences” an object through echolocation. The traditional framing asks: “What is the raw phenomenal quality of echolocation? What does it *feel like*?”

The question is malformed. A bat’s “experience” of an object via echolocation is constituted by integration across multiple levels:

Biological processing: The bat’s nervous system analyzes echo delays, Doppler shifts, and amplitude modulations, producing activation patterns across auditory cortex, superior colliculus, and association areas.

Core affect: The object triggers approach (prey) or avoidance (obstacle) based on threat/reward value, generating valenced arousal states—pleasant low-arousal for a safe roost-mate, unpleasant high-arousal for a predator.

Schema integration (this is where bat consciousness distinctively emerges):

- **Body schema:** The bat’s proprioceptive model integrates flight mechanics with echo feedback. The experience includes where the bat’s body sits relative to the object, what wing adjustments are needed, whether the distance falls within capture range. This is not cognitive calculation but pre-reflective sensorimotor organization.
- **Spatial schema:** Echolocation creates allocentric cognitive maps. The bat experiences objects not as isolated entities but as locations within navigable 3D space—“roost entrance is above and behind me,” “moth is moving left-forward at closing speed.” The spatial schema organizes this into actionable structure.
- **Attention schema:** The bat models its own attention directed at the moth—not thinking “I am aware of the moth” but maintaining an internal model of attention-directedness that shapes subsequent processing. Crucially, this same schema tracks where roost-mates direct their attention, enabling social coordination.
- **Self-schema:** The bat has a sense of itself as the subject to whom these experiences belong, as the agent who will act, as the entity that continues through time. When the bat remembers catching similar prey yesterday, the self-schema provides the continuity: the same “I” experienced that past success.
- **Affective-homeostatic schema:** Ongoing monitoring of hunger (depleted energy stores), fatigue (need for rest), satiation (successful feeding). This provides the motivational context—why this moth matters right now.

Conceptual categorization: The bat has learned categories through experience: “moth-shaped-thing” (prey), “wall” (roost boundary), “other-bat” (social conspecific). These are not linguistic concepts but behavioral categories—patterns of similarity that predict outcomes. The echo pattern gets categorized, activating associated knowledge and action tendencies.

Prediction: The bat predicts the object’s path, prepares motor commands, and anticipates the sensory consequences of action. This creates an integrated forward model: “moth, moving left, will be at location X in 100ms, prepare capture maneuver, expect taste/satiation upon success.”

Phenomenology: The bat's "experience" IS this integrated prediction structured through schemas and categorized by learned concepts. Not raw echo-feels, but meaningful integrated structure: "prey-to-catch-now" organized through spatial navigation, body positioning, attention focusing, self-as-hunter, and homeostatic need.

The phenomenology differs from human vision-based experience not because bat-qualia differ from human-qualia, but because the schematic organization differs:

Structural differences:

- Body schema: Flight-and-echolocation integration vs. bipedal-and-vision integration.
- Spatial schema: Volumetric 3D emphasis vs. surface-based 2D emphasis.
- Attention schema: Simpler social modeling vs. sophisticated theory-of-mind.
- Temporal schema: Present-focused with limited episodic memory vs. extended autobiographical narrative.

Conceptual differences:

- Simpler learned categories (dozens to hundreds vs. thousands of concepts).
- No linguistic elaboration (no words to create fine-grained distinctions).
- Less abstraction (more tied to immediate sensorimotor contexts).

We can understand these differences functionally without accessing raw bat-feels. The question "What is echolocation like?" becomes "How does echolocation structure bat schemas differently from how vision structures human schemas?"

That question is empirically tractable. We study:

- Bat neural organization (revealing schema structure).
- Bat behavioral discriminations (revealing learned categories).
- Bat navigation and hunting (revealing predictive models).
- Bat social behavior (revealing attention schema sophistication).

The mystery is not metaphysical (ineffable raw qualia) but empirical (structural organization we are still mapping).

VIII. Philosophical Implications and Objections

Implication 1: The Hard Problem Reframed Through Schemas

David Chalmers' (1995) "hard problem" asks why physical processes are accompanied by experience. Meaningful integration theory fundamentally transforms this question. I have explained *what* feelings are (meaningful integrations) and *how* they work (predictive categorization of core affect through schemas), but not *why* these processes should feel like anything.

The hard problem transforms, though, once the role of core schemas is recognized:

Old version: Why do neural firings produce raw qualia? **Intermediate version:** Why does meaningful integration generate felt experience? **New version:** Why does schema-organized prediction feel like anything from inside the system?

The new version is more tractable because schemas reveal consciousness as an evolved solution to specific functional problems:

1. Schemas solve coordination problems: An organism with a body needs to coordinate limb movements, predict sensory consequences of action, and maintain postural control. The body schema solves this. But solving it effectively requires the organism to model its body continuously. When this model is accessible to action-selection systems, we call that "feeling embodied."

2. Schemas enable social navigation: A social organism needs to predict conspecifics' behavior, coordinate joint attention, and anticipate others' responses. The attention schema solves this by modeling attention, both others' and one's own. When the self-directed version becomes accessible, we call that "awareness."

3. Schemas create unified agency: An organism acting in the world needs to integrate information across time, maintain continuity of identity, and track outcomes of past actions. The self-schema solves this. When this integration is accessible, we call that "subjectivity."

4. Schemas ground meaning: An organism navigating environments needs to organize knowledge spatially, temporally, and relationally. Spatial and temporal schemas provide this structure. When these organizing frameworks are accessible, they constitute the structure of experience.

The key insight: **Schemas aren't mysterious additions to consciousness but functional solutions that consciousness is.** The body schema doesn't produce a separate feeling of embodiment. *Being accessible to action-selection while modeling the body* just is what embodiment feels like from inside.

Evolutionary gradualism: This makes consciousness less metaphysically special:

- Simple organisms have rudimentary models (chemotaxis in bacteria).
- These models become more sophisticated (sensorimotor maps in invertebrates).
- Models become self-referential (body schema in vertebrates).
- Self-models become accessible to higher-order processing (mammals).
- Models incorporate social prediction (primates).
- Linguistic elaboration enables meta-representation (humans).

At each stage, more sophisticated modeling creates richer phenomenology. There's no sudden leap from unconscious processing to mysterious raw qualia. Instead, schemas that are increasingly accessible to action-selection, planning, and learning systems develop gradually.

Why it feels like something: The remaining mystery (why schema-organized integration should feel rather than occurring "in the dark") may have no deeper answer than: this is what accessible self-models are. Just as digestion is what digestive systems do, and vision is what visual systems do, felt experience is

what accessible self-organizing schemas do. The "why" may resist further explanation because phenomenology is not something added to functional organization but is functional organization experienced from within.

The hard problem is not merely narrowed but dissolved. The question transforms from "why does anything physical create anything mental?" to "why are accessible self-models accompanied by phenomenology?" The RSP answer is that the question is malformed. Under the constitutive identity view, phenomenology is not "accompanied by" self-modeling. It IS what accessible recursive self-modeling is, experienced from within. Asking why self-modeling feels like something is like asking why digestion digests: it mistakes a constitutive relationship for a causal one.

Implication 2: Consciousness as Software

Meaningful integration is a software-level phenomenon, a pattern of organization implemented in biological hardware but not reducible to that hardware. This has several consequences:

Substrate Independence (Partial): In principle, non-biological substrates could implement meaningful integration if they could:

- Generate valenced arousal states (synthetic core affect).
- Learn and apply conceptual categories.
- Build predictive models integrating multiple timescales.
- Implement self-other modeling.

Current AI systems lack these capacities, but they're not obviously impossible.

Multiple Levels of Description: Consciousness is real and causally efficacious at its own level of description, just as software is. Explaining consciousness in purely neural terms would be like explaining Microsoft Word in terms of transistor states: technically complete but missing the explanatory level that matters.

Evolutionary Gradualism: Software can be incrementally improved. This fits with consciousness emerging gradually across evolution and development, rather than appearing suddenly and fully formed.

Objection 1: "Surely I Access Raw Experience!"

The objection: Introspection reveals raw phenomenal qualities. When I see red, I directly apprehend a pure phenomenal redness before any conceptualization. Your theory denies what is most obvious.

Reply: Introspection is not transparent to the structure of experience. What seems like direct access to raw experience is itself a cognitive process, one that presents its output as "raw feel" even though the experience is already conceptually structured.

Consider: When you introspect "red," you access:

- Visual experience (already interpreted: "color," "surface property").
- Emotion associations (warmth, danger, passion, already conceptual).

- Semantic knowledge (roses are red, blood is red, already linguistic).
- Contextual meaning (this red apple looks delicious, already functional).

The seeming "rawness" is the speed and fluency of integration. Conceptualization happens so rapidly that experience feels immediate. But immediacy of access doesn't imply absence of processing.

Analogy: You "directly experience" understanding language, as if meaning flows immediately from sounds. But psycholinguistics reveals extensive unconscious processing (phonological parsing, syntactic analysis, semantic integration). The experience of immediacy doesn't reveal absence of mechanism.

Objection 2: "This Eliminates Real Consciousness"

The objection: By reducing feelings to "mere" integrations of concepts and core affect, you've eliminated the phenomenon we care about: real, felt, subjective experience. You're a crypto-eliminativist.

Reply: This objection presupposes that only raw qualia could be "real consciousness." But meaningful integration is not less real than hypothetical raw qualia. It's more real because it actually exists.

Feelings are not "mere" integrations any more than meaning is "mere" computation or life is "mere" chemistry. The integration is what consciousness *is*. It's felt, it's real, it shapes behavior, it develops, it varies, it matters. Nothing is eliminated except the confused notion of meaning-independent experience.

Consider: Is a symphony "merely" organized sound waves? Is a poem "merely" words on a page? The word "mere" suggests that organization doesn't matter, but organization is precisely what matters. The integration of sounds creates music; the integration of words creates meaning; the integration of core affect with concepts creates felt experience.

Consciousness is not diminished by being understood as integration. It's clarified.

Objection 3: "This Makes Consciousness Too Cognitive"

The objection: You emphasize conceptual categorization and now schemas, but infants and animals have feelings without sophisticated concepts or explicit self-models. Feelings seem more primitive and immediate than cognitive processing or structural organization.

Reply: This objection misunderstands the role of both concepts and schemas in meaningful integration. Neither concepts nor schemas need be linguistic, deliberate, or sophisticated. They're patterns of organization that structure experience.

Even infants and animals have proto-concepts and foundational schemas:

Proto-concepts:

- Perceptual categories (prey vs. non-prey, approaching vs. receding).
- Affordance categories (edible, dangerous, comfortable).
- Social categories (mother, stranger, threat).

Core schemas:

- Body schema develops through sensorimotor experience (even neonates have rudimentary body awareness).
- Spatial schema emerges as infants navigate environments.
- Self-schema begins forming around 18-24 months but has precursors.
- Attention schema operates even before explicit self-awareness.
- Affective-homeostatic schema is present from birth.

These needn't be linguistic or explicit. The body schema is a sensorimotor organization: a set of predictions about how the body moves and what it can do. The attention schema is a forward model of attention, not a thought about attention. When core affect is organized through these schemas and categorized using proto-concepts, even simple organisms have feelings (though simpler than adult human feelings).

The sophistication of consciousness scales with both schematic complexity and conceptual sophistication:

- Simple organisms: basic valenced responses organized through minimal schemas.
- Mammals: more differentiated feelings based on richer schematic structure and categorization.
- Primates: self-other modeling enables complex social emotions organized through sophisticated attention schemas.
- Humans: linguistic concepts enable extraordinary emotional granularity built on the shared mammalian schematic foundation.

This gradation is a feature, not a bug. It explains why consciousness isn't all-or-nothing but varies across development and evolution: both schemas and concepts develop, creating increasingly sophisticated meaningful integration.

Objection 4: "Schemas Are Just Structure, Not Experience"

The objection: You've added another layer (core schemas) but this just pushes the problem back. Schemas are structural organizations, not phenomenology. You still haven't explained what-it's-like, just added more mechanism.

Reply: This objection assumes phenomenology must be something *over and above* structure and mechanism, precisely the raw qualia assumption I am rejecting.

Schemas are not "just" structure any more than music is "just" organized sound or meaning is "just" computational relationships. The organization *is* what creates the phenomenon. When affective states are organized through body schema, spatial schema, self-schema, and attention schema, the result *is* phenomenology. Not something added to structure, but the structure itself experienced from within.

Consider: Your visual experience of depth isn't something added to binocular disparity. It's how binocular disparity appears from the inside. Your experience of agency isn't added to motor preparation. It's how motor preparation structured through the self-schema feels. Your awareness of attending isn't separate from attention. It's attention structured through the attention schema and thereby made accessible to reflection.

The schemas are transparent: they present themselves as reality rather than as models. You don't experience "my body schema is representing my arm at location X"; you experience "my arm is at location X." This transparency creates the sense of immediacy, the feeling that we access raw experience rather than structured predictions.

The crucial insight: **phenomenology is not an added extra but is what organized, integrated processing feels like from inside the system.** Schemas don't produce separate phenomenal properties. They *are* the organization that constitutes phenomenology when integrated with core affect and conceptual categorization.

This is why damage to schema-generating regions (temporoparietal junction for attention schema, parietal cortex for body/spatial schemas) disrupts consciousness itself, not just unconscious processing. The schemas aren't mechanisms that somehow produce separate phenomenal feels. They're the structural organization that, when operating, *is* felt experience.

Objection 5: "Social Prediction Doesn't Explain Why It Feels Like Something"

The objection: You've explained that consciousness involves recursive self-modeling for social prediction. But this still doesn't explain the hard problem. Why should recursive modeling feel like anything at all? You could have all this modeling occurring "in the dark" without phenomenology.

Reply: This objection reveals a persistent confusion about what phenomenology is. It assumes phenomenology is something separable from recursive self-modeling, an extra property that could be present or absent while leaving the functional structure intact.

But consider: Can you have a whirlpool without a center? Can you have music without temporal patterns? Can you have meaning without relationships between symbols? In each case, the supposed "extra" property is not separate from the structure. It *IS* the structure from a particular perspective.

Phenomenology is not an added extra to recursive self-modeling. It's what recursive self-modeling is like from inside the system executing it. When a system:

- Models its own states (self-representation).
- Uses these models predictively (anticipating future states).
- Feeds predictions back into itself (recursion).
- Does this in service of social navigation (modeling self-as-perceived-by-others).
- Has this process accessible to action-selection, learning, and planning.

...then that system, by virtue of this structure, has phenomenology. The feeling is not something added. It's intrinsic to being the system executing these recursive loops.

The zombie argument collapses: The philosophical zombie thought experiment asks us to imagine a being physically identical to us but lacking phenomenology. But if phenomenology is intrinsic to recursive self-modeling (not a separate addable property), then:

- A being with our recursive architecture would have phenomenology by virtue of that architecture.

- A being lacking phenomenology would necessarily have different architecture (no recursion, or unconscious self-monitoring).
- The zombie scenario is incoherent, like asking for a whirlpool with identical water dynamics but no center.

The conceivability argument (Chalmers, 2010) claims that zombie worlds are conceivable because phenomenal concepts have distinct primary and secondary intensions. But if phenomenal concepts are themselves constituted by recursive self-modeling — as RSP claims — then the primary intension of a phenomenal concept IS the self-modeling state. A zombie, lacking the recursive architecture, would lack not just the phenomenology but the very concepts needed to conceive of the zombie scenario. This is not a merely verbal point: it means the conceivability intuition itself depends on having the architecture whose absence defines the zombie. Conceivability, in this case, does not track metaphysical possibility.

The transcendental reading of RSP (§III) dissolves the zombie argument at the root. Assume the zombie has Level 3b intact — recursive self-modeling, social prediction, the strange loop fully functional. Under RSP, Level 3b is where the subject position comes into existence: the "I" is constituted by the recursive loop, not added to it. If Level 3b is functional, the zombie has a subject position and is therefore conscious. The zombie with Level 3b intact is not a zombie. If we insist that Level 3b can operate without a subject position, we are positing something additional to the functional architecture — but then that additional thing becomes an empirical target rather than a metaphysical mystery, and the hard problem is converted from a conceptual impossibility into a testable hypothesis about what, specifically, the architecture is missing. Either way, the zombie argument loses its force: it either fails (because Level 3b entails the subject position) or becomes an empirical question (about what beyond Level 3b might be needed).

It is worth noting that RSP's approach to the zombie problem has a close relative in Carruthers' (2000) constitutive higher-order thought theory, which argues that phenomenal consciousness is constituted by (not merely accompanied by) higher-order representations. RSP extends Carruthers' insight in two directions. First, it specifies the *content* of the higher-order representations: they are not generic "thoughts about experiences" but predictions generated by social-modeling machinery applied to self. Second, it grounds the higher-order architecture in a developmental and evolutionary account: the recursive self-modeling capacity emerges from the social prediction machinery that evolved for navigating group life, giving the higher-order representations a specific origin and function that Carruthers' more abstract account leaves open.

RSP also addresses what Chalmers (2018) calls the "meta-problem" of consciousness: why do we have the intuition that consciousness is hard to explain, even if it is in fact functional? Under RSP, the answer is straightforward: the recursive self-model (Level 3b) represents its own processing as having intrinsic phenomenal character, because that is how the self-model must represent states it is modeling from the inside. The "hardness" intuition is itself a prediction of the architecture — a system that models its own experience will inevitably represent that experience as having properties that resist third-person description, because first-person access to the self-model is structurally different from third-person description of the same process.

Why it feels like something specific: Different feelings (anger vs. joy vs. fear) arise from different meaningful integrations, different combinations of core affect, schema activation, conceptual categorization, and social-predictive context. The "what-it's-like" of anger is the specific pattern of:

- Unpleasant high-arousal affect.
- Body schema primed for approach/aggression.
- Attention schema focused on offender.
- Self-schema as wronged party.
- Social prediction of confrontation consequences.
- Recursively modeling yourself being angry, others perceiving your anger, yourself perceiving their perception...

This specific pattern of integration and prediction CONSTITUTES what anger feels like. There's no additional "anger-qualia" floating independently. The phenomenology is the pattern experienced from within.

Objection 6: "This Makes Non-Social Animals Unconscious"

The objection: Your account emphasizes social prediction so heavily that it seems to imply non-social animals (solitary mammals, reptiles) lack consciousness. But surely a solitary cat has experiences even without engaging in social prediction.

Reply: The social-predictive account doesn't require constant social interaction for consciousness. It claims that the CAPACITY for self-modeling in social contexts creates the architecture enabling phenomenology.

Three levels of sociality in consciousness:

1. Core affect without social modeling (fish, reptiles, simple mammals):

- Have valenced arousal states.
- Have basic schemas (body schema, spatial schema).
- Experience approach/avoidance, pleasure/pain.
- BUT lack sophisticated self-models and recursive loops.
- Result: Basic affective experience without self-consciousness.

2. Self-modeling without rich theory of mind (solitary mammals like cats):

- Have core affect and schemas.
- Have self-models (recognize self vs. other, have body ownership).
- Can predict own immediate future states.
- BUT limited theory of mind (don't model others' mental states richly).
- Result: Self-conscious experience but simpler than social species.

3. Full recursive social-prediction (social mammals, especially humans):

- Have all of the above PLUS.
- Rich theory of mind (model others' beliefs, desires, intentions).
- Recursive self-modeling (model self-as-perceived-by-others).
- Cultural concepts enabling fine-grained elaboration.
- Result: Rich, linguistically-elaborated, socially-embedded consciousness.

Even solitary mammals retain the neural architecture for self-modeling that evolved originally for social purposes. A cat alone still uses self-predictive modeling to anticipate its own hunger, fatigue, interest in prey, and satisfaction from catching. This self-model is simpler than human recursive social-prediction but still enables phenomenology richer than mere affective response.

The claim is not "only social interactions create consciousness" but "the capacity for self-modeling in social contexts creates the neural architecture that enables phenomenology, even when not currently engaged socially."

Apparent counterexamples: meditation, flow states, and dreams. Three forms of consciousness seem to involve *less* social modeling, not more, creating potential difficulties for the social-predictive account:

Meditation: Long-term practitioners report dissolving the self-model, experiencing consciousness without a strong sense of "I." Under the RSP framework, this is not the absence of Level 3 but its systematic *attenuation*: the recursive loop still operates but with reduced gain (lower precision weighting on self-referential predictions). The meditator does not lose Level 3; they modulate it. That meditation requires years of practice to achieve this modulation is itself evidence that the default mode is recursive self-modeling. The architecture must be actively quieted.

Flow states: Absorption in a task with loss of self-consciousness. Under RSP, flow represents a shift in precision allocation: Level 2 schemas (particularly the body schema and attention schema engaged with the task) receive maximal precision, while Level 3 recursive self-modeling receives minimal precision. The self-model doesn't vanish. The person in flow can be startled out of it by social interruption, instantly restoring self-awareness. But it operates at reduced gain. Consciousness persists (the person is vividly aware) but self-*consciousness* is temporarily attenuated.

Dreams: Dream consciousness often includes vivid phenomenology without active social interaction. But dreams frequently involve *simulated* social scenarios (other people, social situations, emotional encounters), which is exactly what Level 3's social prediction machinery would produce when running unconstrained by external input. The bizarre character of dreams, with their violations of physics, identity confusion, and impossible scenarios, is consistent with Level 3 operating without the error-correcting feedback from real social interaction.

In each case, the RSP framework accommodates the phenomenon through precision modulation rather than architectural absence: the three-level architecture remains, but the relative weighting of levels shifts.

A sharper version of this objection comes from Merker (2007), who argues that decerebrate infants — lacking the entire neocortex and therefore all cortical social machinery — display purposeful behavior

suggesting some form of consciousness. If RSP requires Level 3 recursive social prediction for phenomenal consciousness, and decerebrate infants lack the neural substrate for Level 3, then RSP must either deny these infants have any form of consciousness or revise its architecture.

RSP's response is that the framework already accommodates this case through its graduated consciousness model. Decerebrate infants would have Level 1 core affect (brainstem and hypothalamic circuits are intact) and possibly rudimentary Level 2 schemas (some subcortical body-schema processing survives). What they lack is Level 3: recursive self-modeling and conceptual categorization. Under RSP, they have valenced experience (things feel good or bad) and possibly rudimentary structured experience (body-awareness, spatial orientation) but not full phenomenal consciousness with self-awareness and conceptual content. This is not a denial of their experience; it is a specific prediction about its character: affective, pre-reflective, and non-self-aware. The prediction is consistent with the behavioral evidence Merker cites (approach/avoidance, emotional expression, pain responses) while denying the richer phenomenology that requires recursive social prediction.

The Social-Centricity Qualification

The claim that consciousness is "fundamentally social" requires clarification. RSP does not claim that social interaction is necessary for consciousness at every moment. Meditation, solitary reflection, and dreams all involve consciousness without active social engagement. The claim is architectural: the neural machinery that constitutes Level 3b (recursive self-modeling) evolved for social prediction and retains its social-predictive structure even when deployed non-socially. A meditator's self-awareness uses the same TPJ/STS circuits that model others' mental states. The sociality is in the architecture's origin and structure, not in its moment-to-moment deployment.

Objection 7: "Infinite Regress in Recursive Modeling"

The objection: If consciousness is recursive self-modeling, don't we have infinite regress? If I'm modeling myself modeling myself modeling myself... where does it stop? Don't we need a base level that isn't itself modeled?

Reply: Yes, and there is one: core affect and basic schemas. The recursion doesn't go down infinitely. It has a grounding:

Base level: Core affect (valence/arousal) organized through basic schemas (body schema, spatial schema). This level is experienced but not self-modeled. The infant feeling hunger doesn't model itself feeling hunger. It simply feels hungry.

First-order self-model: The brain models its own states. "I am hungry, I am in my room, I am reaching for bottle." This creates basic self-consciousness.

Second-order recursive model: The brain models itself modeling states. "I am aware that I am hungry." This is what we typically call consciousness: the ability to think about your thinking.

Third-order and beyond: "I am aware that I am aware that I am hungry" and so on. But in practice, the recursion is shallow, usually only 2-3 levels deep in ordinary experience. You don't need infinite recursion,

just enough to create predictive self-models.

Why recursion is bounded:

- Cognitive resources are limited (can't maintain infinite loops).
- Functional purposes are satisfied by shallow recursion (2-3 levels sufficient for social prediction).
- The loop naturally stabilizes at depth where prediction error minimization works well.

The "strange loop" isn't literally infinite. It's a few levels of recursion creating self-sustaining patterns. Hofstadter's point wasn't that consciousness requires infinite levels but that it requires closed loops, with level N feeding back to influence level N-1, creating self-reference.

Objection 8: "This Is Still Just Functionalism"

The objection: You've given a sophisticated functional story about recursive self-modeling, but ultimately this is just functionalism, defining consciousness by what it does. Functionalism faces the classic objection: couldn't something implement these functions without phenomenology?

Reply: This objection assumes phenomenology is something separable from function. The social-predictive account rejects this assumption. Phenomenology is not a separate property that might or might not accompany function. It IS what certain functions are like from inside.

Not generic functionalism but specific architectural functionalism: The claim is not "any system with the right inputs/outputs is conscious" but "systems with recursive self-modeling architecture enabling social prediction have phenomenology intrinsically."

The architecture matters specifically:

- Must have valenced states (not just information processing).
- Must have self-models (not just external models).
- Must use self-models predictively (not just descriptively).
- Must have recursion (self-models influencing themselves).
- Must be accessible to action-selection (not unconscious monitoring).

Systems lacking these specific features wouldn't just lack phenomenology. They'd lack the functions themselves. You can't have recursive self-modeling for social prediction without phenomenology any more than you can have a whirlpool without circular motion.

Different from classic functionalism: Classic functionalism defines mental states by causal role (belief, desire, pain) and faces the objection that zombies could have those roles without experience. The social-predictive account defines consciousness by architectural structure (recursive self-models) and argues that phenomenology is intrinsic to that structure when experienced from within.

The zombie objection doesn't work because a zombie with identical recursive architecture would have phenomenology. A zombie truly lacking phenomenology would have different architecture. The intuition that zombies are possible rests on wrongly imagining phenomenology as separable from recursive structure.

Why Level 3b, Not Level 2?

A natural objection: if schemas organize experience at Level 2, why isn't Level 2 sufficient for phenomenal consciousness? The answer requires a distinction that earlier formulations of RSP left implicit: phenomenal consciousness comes in degrees, and what Level 3b adds is not phenomenality as such but the *recursive, self-aware dimension* of phenomenality.

Level 1-2 phenomenality is genuine. There IS something it is like to be a creature with organized body-awareness, spatial orientation, and affective-homeostatic regulation, even without recursive self-modeling. A fish with Level 1 core affect has valenced experience: things feel good or bad. A mammal with Level 2 schemas has richer structured experience: its body schema gives it felt embodiment, its spatial schema gives it felt orientation, its affective-homeostatic schema gives it felt hunger and satiation. These are not "mere information processing"; they are forms of phenomenality, simpler than adult human consciousness but genuine nonetheless.

What Level 3b adds is the *recursive dimension*: a model of the system as a whole, viewed from multiple social perspectives, recursively predicting its own future states. This recursive self-model transforms experience from something the system *undergoes* into something the system *knows it undergoes*. The distinction is between phenomenality (which admits of degrees across Levels 1-2-3) and self-aware phenomenality (which requires the recursive architecture of Level 3b). A thermostat has neither. A fish has basic valenced phenomenality. A cat has schematically structured phenomenality. An adult human with Level 3b has self-conscious phenomenality — experience that includes a model of the experiencer within the experience itself.

The developmental evidence supports a graduated emergence. Neonates with rudimentary Level 1-2 processing have genuine affective experience: they feel pain, comfort, and hunger. Toddlers with developing schemas have richer structured experience: felt embodiment, spatial awareness, and proto-emotional states organized through body and affective-homeostatic schemas. The 18-24 month mirror self-recognition milestone marks the emergence of a rudimentary self-schema at the upper boundary of Level 2. Between ages 2 and 5, implicit precursors of Level 3b (perspective-taking in pretend play, emerging theory of mind, early forms of self-conscious emotion such as embarrassment and pride) provide increasingly self-aware experience. Full recursive self-modeling (Level 3b) at 4-5 years adds the strange loop that constitutes the kind of self-conscious phenomenality distinctive of mature human consciousness. The transition is gradual, not a threshold.

The constitutive identity claim, refined. The claim that phenomenology IS recursive self-modeling applies specifically to the self-aware, self-referential dimension of consciousness distinctive of mature human experience. It does not deny that simpler forms of phenomenality exist at Levels 1-2. Fish have valenced experience. Toddlers have rich schematic experience. What they lack is the recursive dimension, and it is this recursive dimension, not phenomenality as such, that the dissolution strategy targets. The hard problem, as traditionally formulated, asks why *self-aware* experience exists: why there is something it is like *for me* to see red, not merely why chromatic processing occurs. RSP's answer — that the "for me"

is constituted by recursive self-modeling — dissolves this question without denying that simpler phenomenality exists at lower architectural levels.

This graduated picture preserves what is right about the original sharp formulation (Level 3b is genuinely distinctive and constitutively important) while avoiding its false implication (that creatures without Level 3b have no phenomenality at all). The difference between a system that processes information, a system that has structured experience, and a system that models itself having structured experience is real and principled. But it is a difference of degree and kind within phenomenality, not a boundary between phenomenality and its absence.

Cultural Variation

The cultural variation objection, namely that different cultures having different emotion concepts threatens universality, is addressed in §VI. Under RSP, cultural variation in phenomenology is a central prediction, not a problem. Different conceptual repertoires (Level 3a) produce genuinely different experiences from the same core affect (Level 1), exactly as the architecture predicts.

Objection 9: "What About Mary? The Knowledge Argument"

The objection: Jackson's (1982) Mary knows all physical facts about color vision but has never seen red. When she finally sees red, she learns something new: what it's *like* to see red. This shows that physical/functional knowledge doesn't capture phenomenal character. Your RSP model, however sophisticated, is still physical/functional knowledge. Mary could know the entire RSP architecture and still not know what red looks like.

Reply: The knowledge argument assumes that "knowing all physical facts" and "having the experience" are separable. Under the RSP framework, they are not, but the reason is subtle.

Mary in her black-and-white room lacks something specific: she has never run the Level 1-2-3 integration process for red. She knows the architecture (she can describe it), but she has never *instantiated* it for this particular input. When she sees red for the first time, her brain performs a meaningful integration she has never performed before: interoceptive response to red light (Level 1), schema-organized spatial-chromatic experience (Level 2), and self-model updated to include "I am now seeing red" (Level 3). This is genuinely new. Not because some ineffable quale was missing from her physical knowledge, but because *running* a computational process is different from *describing* it.

The analogy: knowing all the rules of chess is different from playing a particular game. The experience of playing (the specific integration of perception, emotion, strategy, and self-awareness in that moment) is not a separate metaphysical substance; it is the process of meaningful integration occurring in real time. Mary gains new *integration*, not new *substance*.

This response is available to any sophisticated functionalist (Dennett, 1991; Churchland, 1985). What RSP adds is specificity: I can say exactly *which* integration Mary lacks (the three-level recursive processing of chromatic input through her particular schemas and self-model) and predict that her experience of red will

be shaped by her conceptual repertoire. A vision scientist's experience of red is richer than a layperson's, because her Level 3a conceptual categories are more differentiated.

The knowledge argument succeeds in showing that propositional knowledge ("red has wavelength 700nm") differs from experiential knowledge ("this is what red looks like"). RSP agrees and explains the difference: propositional knowledge is Level 3a conceptual content; experiential knowledge is the full three-level integration. The gap between them is real but functional, not metaphysical.

Objection 10: "The Cogito: Conscious Experience Is the One Thing We Cannot Doubt"

The objection: Descartes (1641) argued that conscious experience is the one phenomenon whose existence cannot coherently be doubted, because the act of doubting is itself an experience. Some argue that the RSP architecture could in principle run without any experience at all (the zombie scenario), yet no one can coherently doubt that they are experiencing right now. The Cogito reveals something about consciousness that no third-person functional account can capture. A dissolution strategy, however sophisticated, still describes the machinery from the outside. It does not explain why there is an inside.

Reply: The reply must begin where it matters most: with the constitutive identity thesis. RSP does not claim that the recursive architecture *produces* experience as an output, the way a factory produces goods. RSP claims that the recursive self-modeling process at Level 3b is the experience, viewed from within. The dissolution strategy does not describe machinery from outside; it identifies the inside view with the machinery. There is no gap between "the architecture running" and "there being something it is like," because the running of the self-model at Level 3b is the something-it-is-like.

From this constitutive identity, the Cogito follows as a prediction rather than a challenge. Consider what happens when a system with a functioning Level 3b asks "am I experiencing anything?" The system must consult its own recursive representation of its own states. That representation necessarily reports experience, because the representing and the experiencing are the same process. A running program that queries whether any program is running on its processor cannot return "no." The self-referential structure of the query guarantees an affirmative answer, not because the system is confused, but because the querying is an instance of the very process being queried. And because the constitutive identity holds, this is not merely an internal state update that happens to return "yes." The running of the query IS phenomenal experience, not a report about phenomenal experience. The first-person perspective just is what instantiation of the recursive self-model looks like from within.

RSP therefore predicts that any system with a stable recursive self-model (Level 3b; see Appendix D, §D.8 on the self as state observer) will genuinely have, not merely report, Cogito-like certainty about its own experience. The certainty Descartes identified is real. But it is a structural consequence of recursive self-modeling, not evidence of a non-physical substance. Descartes was right that we cannot doubt our own experience. He was wrong about why. The indubitability follows from the constitutive identity between recursive self-modeling and phenomenology, not from the existence of a non-physical *res cogitans*.

A residual question remains: why should we accept the constitutive identity claim rather than treating it as a relabeling of the hard problem? RSP's answer is that constitutive identity is a primitive of the theory, in the same way that the relationship between electromagnetic fields and light is a primitive of Maxwell's equations, not itself derived from something deeper. The claim is not that RSP *derives* why recursive self-modeling feels like something. The claim is that once we identify phenomenology with recursive self-modeling (as §IV argues on independent grounds), the Cogito becomes a prediction rather than a mystery. The identification does real work: it generates falsifiable predictions (§IX), explains clinical dissociations (Appendix G), and unifies five mathematical formalisms under a single master functional (Appendix E). An identification that does this much explanatory work earns its primitive status, just as Maxwell's identification of light with electromagnetic radiation earned its status by predicting the speed of light from first principles (Prediction 4, §IX, provides the analogous empirical anchor for RSP: recursive depth $D_{\max}$ should correlate with metacognitive capacity across species).

Three consequences follow. First, a system without Level 3b (an insect, a thermostat) lacks the recursive structure that generates Cogito-like certainty. It does not doubt its experience, but neither does it affirm it. It simply is not the kind of system for which the question arises. Second, the Cogito is species-relative in its depth: a great ape with limited recursive depth (

$$D_{\max} \approx 1-2$$

) has a correspondingly limited form of self-certainty, measurable via metacognitive monitoring paradigms (Hampton, 2001; Call & Tomasello, 2008). Third, an artificial system that instantiated Level 3b with convergent recursive dynamics would generate its own Cogito, not because someone programmed it to report experience, but because the recursive self-model would make its own phenomenality self-evident from within.

IX. Empirical Implications and Predictions

A theory of consciousness that generates no testable predictions is not a scientific theory. The RSP model, grounded in the three-level architecture and formalized across five appendixes, makes seven specific predictions that could be confirmed or refuted by existing experimental methods.

Prediction 1: Emotional Granularity Interventions

If concepts constitute feelings (not just label them), then teaching people new emotion concepts should create new phenomenology, not just new vocabulary.

Testable prediction: Adults taught fine-grained emotion distinctions (teaching the difference between "anxious," "tense," "worried," "apprehensive") should:

- Report qualitatively different experiences in situations previously labeled generically.
- Show neural differentiation in emotion-relevant brain regions.
- Demonstrate improved emotion regulation (because finer concepts enable more precise prediction and intervention).

Evidence: Studies of emotion differentiation support this. Higher emotional granularity predicts better regulation, less psychopathology, and more adaptive responding (Barrett et al., 2001; Kashdan et al., 2015). Critically, interventions that teach emotion concepts improve functioning, suggesting concepts change experience, not just description.

Prediction 2: Cross-Cultural Phenomenology

If cultures with unique emotion concepts create unique phenomenologies, one should find:

- Neural differences corresponding to culturally unique emotions.
- Behavioral differences in emotion regulation strategies.
- Phenomenological reports that don't map onto other languages.
- Developmental paths shaped by concept acquisition.

Evidence: Emerging research finds:

- Culture-specific neural patterns for emotions (Chiao et al., 2008).
- Systematic variation in which emotions are valued/regulated (Tsai, 2007).
- Emotion concepts that lack translation (Wierzbicka, 1999).
- Emotion socialization varies across cultures, predicting different emotional outcomes (Wang & Fivush, 2005).

Prediction 3: Developmental Sequence of Schemas and Concepts

If emotional experience requires both schemas and concepts, then:

- Very young infants should show undifferentiated core affect organized through rudimentary schemas.
- Schema development should precede conceptual elaboration (body schema before emotion words).
- Emotional differentiation should track both schema maturation and concept acquisition.
- Language development should predict emotional granularity beyond what schemas alone enable.
- Social context richness should predict both schema sophistication and conceptual elaboration.

Evidence: Developmental research confirms this layered emergence:

- Neonates show body awareness and affective responses (schema + core affect without concepts).
- Body schema and spatial schema develop 0-6 months through sensorimotor experience.
- Self-schema emerges 18-24 months with mirror self-recognition.
- Attention schema develops through social interaction (joint attention predicts later theory of mind).
- Emotion concepts are acquired 2-5 years, following schema development.
- Emotional granularity increases throughout childhood as both schemas refine and concepts proliferate.

This developmental ordering is formally constrained by the RSP architecture (see Appendixes B–D). In the free energy framework, minimizing F_2 requires stable inputs from F_1 ; in the RL framework, state representation learning (Level 2) requires a functional reward signal (Level 1); in the control-theoretic framework, cascade stability requires inner loops to be commissioned before outer loops ($\tau_1 < \tau_2 < \tau_3$).

The child's unsupervised play (reaching, crawling, pointing, babbling) is the exploration policy that builds Level 2 state representations from Level 1 reward signals, precisely as the unsupervised RL formulation (Appendix C, §C.5) predicts. This makes the developmental sequence a strong prediction of the theory: any organism with this architecture *must* develop its levels in this order.

Testable predictions:

- Children with richer sensorimotor experiences should develop more sophisticated body/spatial schemas, enabling finer affective differentiations even before acquiring emotion words.
- Theory-of-mind development should correlate with attention schema maturation (measured through perspective-taking tasks) more than with language ability alone.
- Interventions targeting schema development (movement therapy, spatial navigation training) should enhance emotional awareness even without explicit emotion teaching.

Prediction 4: Schema Damage and Consciousness Deficits

If schemas constitute the structural organization of consciousness, then damage to schema-generating regions should produce specific consciousness deficits:

Testable predictions:

- Damage to temporoparietal junction (attention schema) should disrupt not just attention but subjective awareness itself (confirmed: hemispatial neglect involves consciousness deficit).
- Damage to parietal cortex (body/spatial schemas) should alter phenomenology of embodiment and spatial experience (confirmed: out-of-body experiences, body ownership disruptions).
- Damage to midline cortical structures (self-schema) should affect sense of continuity and identity (confirmed: various dissociative experiences).
- Crucially, these deficits should be more severe than damage to areas with equivalent functional importance but not involved in schema generation.

Evidence: Neuropsychology supports schema-specific deficits:

- TPJ stimulation induces out-of-body experiences (Blanke et al., 2002).
- Right hemisphere TPJ damage causes hemispatial neglect: patients lack awareness of left space entirely.
- Parietal damage disrupts body ownership, as in somatoparaphrenia (denial of limb ownership following right parietal damage).
- These are not mere processing deficits. They are disruptions of conscious experience itself.

Prediction 5: Therapeutic Mechanisms

If emotional experience involves meaningful integration, therapy should work by:

- Teaching new conceptual categories (cognitive restructuring).
- Changing predictions about situations (exposure therapy).

- Updating self-models (psychodynamic therapy).
- Modifying core affect regulation (somatic therapies).

Each intervention targets a different component of meaningful integration.

Evidence: Different therapies work through different mechanisms but all involve changing how experiences are integrated:

- CBT teaches reconceptualization (Beck, 1979).
- Exposure therapy updates predictions (Craske et al., 2014).
- Emotion-focused therapy differentiates feelings (Greenberg, 2002).
- Acceptance-based approaches change relationship to experience (Hayes et al., 2006; see also Kabat-Zinn, 1990 for mindfulness-based interventions).

The meaningful integration framework predicts that effective therapy always involves changing how core affect is interpreted, predicted, or regulated. This is exactly what we observe.

Prediction 6: Clinical Dissociations Should Map the RSP Architecture

If consciousness has the three-level architecture proposed here, then neurological and psychiatric conditions should produce *component-specific* consciousness deficits, not global dimming of awareness but selective disruption of the particular architectural component that is damaged. The RSP model predicts three categories of clinical dissociation, each mapping to a distinct level or interface within the architecture.

Self-Model Failures (Level 3b)

When the recursive self-modeling loop malfunctions, patients should remain conscious but experience specific distortions in the "who am I?" component of the strange loop:

Cotard's delusion: Patients believe they are dead, do not exist, or have lost their organs. The self-model has become so impoverished that it generates the prediction "I do not exist," yet the system continues to process and report on this prediction. The recursive loop still runs but models a non-self, creating the paradox of a conscious being asserting its own non-existence. This is the strange loop turned inside out (Young & Leafhead, 1996; Gerrans, 2000).

Depersonalization/derealization: Patients feel detached from their own mental processes, observing themselves as if from outside. The self-model still operates but is not properly integrated with first-person experience. The recursive loop runs in "spectator mode," generating predictions about the self that are not felt as belonging to the self. The phenomenology shifts from "I think" to "something is thinking" (Sierra & Berrios, 1998; Medford et al., 2005).

Dissociative identity disorder: Multiple self-models coexist within one neural substrate, each with its own recursive loop, predictive patterns, and phenomenological character. DID demonstrates that the self-model is not a fixed property of the brain but a constructed prediction, and that the construction process

can fork. Each identity maintains its own Level 3 strange loop with different conceptual categories and social predictions (Putnam, 1989; Reinders et al., 2003).

Prediction and Agency Failures (Level 2–3 Interface)

When the temporal prediction components break down (the ability to predict one's own future states and attribute agency to one's actions), a different pattern of deficits emerges:

Alien hand syndrome: The hand performs purposeful actions that the patient does not experience as self-initiated. The motor command is generated, but the self-model does not predict it, creating a prediction error so large that agency is denied entirely. The hand belongs to the body schema (Level 2) but is disconnected from the self-model (Level 3). Actions without recursive self-prediction feel alien (Della Sala et al., 1991; Marchetti & Della Sala, 1998).

Anosognosia: Patients deny obvious disabilities such as paralysis after stroke. The self-model cannot update to incorporate new information about the body: the recursive update step is blocked, and the self-model keeps predicting a functional body despite contradictory evidence. This reveals that the self-model is not a passive mirror but an active prediction that can resist disconfirmation (Ramachandran, 1996; Prigatano, 2010).

Phantom limb: Patients feel vivid sensations in amputated limbs. The body schema (Level 2) persists in predicting a limb that no longer exists, and the self-model (Level 3) integrates these predictions into phenomenal experience. Consciousness is generated by the predictive model, not by sensory input (Ramachandran & Hirstein, 1998; Melzack, 1990).

Social Prediction Failures (Level 3a–3b)

Because consciousness in the RSP model is fundamentally social, built from machinery evolved to model other minds, disruptions to social cognition should produce distinctive consciousness alterations:

Capgras delusion: Patients recognize familiar faces but believe the person has been replaced by an identical impostor. The visual recognition system works, but the social-predictive model generates a massive prediction error: "this person looks like my wife but does not feel like my wife." The autonomic affective familiarity signal (Level 1) fails to accompany the perceptual recognition (Level 2), and the self-model (Level 3) confabulates the "impostor" explanation, revealing how failures at lower levels cascade upward through the architecture (Ellis & Young, 1990; Hirstein & Ramachandran, 1997).

Prosopagnosia: The inability to recognize faces disrupts the "other-models" component of the strange loop. Without stable other-models, the "predict self as seen by others" step is impaired. Patients often report changes in social phenomenology: social situations feel different, less rich, more anxious, because the recursive social prediction loop cannot properly engage (Duchaine & Nakayama, 2006; Behrmann et al., 2005).

Autism spectrum differences: Differences in social prediction machinery produce systematically different phenomenology. Individuals on the autism spectrum often show preserved or enhanced Level 1–2 processing (sometimes heightened sensory awareness) with differences in Level 3 recursive social

modeling: difficulty predicting others' mental states and, in turn, differences in the recursive self-model that depends on those other-models. This is not a deficit but a different configuration of the RSP architecture (Baron-Cohen, 1995; Frith, 2003; Lombardo et al., 2010).

The Clinical Logic

The critical point is not that any single clinical condition proves the RSP model, but that the systematic mapping of symptoms to architectural components provides a powerful pattern of converging evidence. Each condition disrupts a *specific* component while leaving others intact:

- A patient with Capgras delusion does not lose all consciousness. They lose the specific social-predictive component the model predicts they should lose.
- A patient with anosognosia is conscious but cannot update one aspect of their self-model.
- A patient with alien hand syndrome has intact self-awareness except for agency over one limb.
- A patient with phantom limb has a body schema that persists without sensory confirmation, demonstrating that consciousness is model-driven rather than input-driven.

This selective impairment is the hallmark of a genuine architectural theory rather than a vague philosophical gesture. Competing theories do not predict this pattern as cleanly: IIT would predict consciousness deficits proportional to reduced Φ (integrated information), with no reason for deficits to sort into the specific categories observed. Global Workspace Theory would predict deficits proportional to reduced broadcasting, not component-specific dissociations. Only the RSP model predicts that clinical disruptions should map onto a specific three-level architecture with identifiable subcomponents, and the clinical evidence confirms this prediction across dozens of documented conditions.

Prediction 7: Social Deprivation Should Selectively Impair Level 3 While Preserving Levels 1-2

If consciousness has the three-level architecture proposed here, with Level 3 (recursive social prediction) requiring social input to develop, then severe social deprivation during critical periods should produce a characteristic dissociation: preserved core affect and basic schemas with impoverished recursive self-modeling, theory of mind, and emotional granularity.

Testable predictions:

- Socially deprived individuals should show intact core affect responses (valence/arousal). **Confirmed:** Genie Wiley displayed fear, pleasure, anger, and attachment (Curtiss, 1977); Romanian orphans showed distress and pleasure responses (Nelson, Fox & Zeanah, 2014).
- They should show functional body and spatial schemas. **Confirmed:** feral children navigate environments, have body awareness, manipulate objects; Genie walked and explored upon rescue.
- They should show impaired or absent theory of mind. **Confirmed:** Genie showed minimal capacity for perspective-taking; institutionalized Romanian children showed delayed theory of mind development (Yagmurcu, Berument & Celimli, 2005).

- They should show reduced emotional granularity. **Confirmed:** severely deprived children across all documented cases show limited emotional vocabulary and difficulty differentiating complex social emotions from basic affects.
- They should show impaired mirror self-recognition and autobiographical memory. **Confirmed:** Romanian orphans showed delayed self-recognition; feral children showed absent or fragmentary autobiographical narrative.
- Earlier social intervention should produce better Level 3 recovery than later intervention. **Confirmed:** the Bucharest Early Intervention Project found that children placed in foster care before age 2 showed significantly greater recovery in social cognition, attachment, and self-modeling than those placed later (Nelson, Fox & Zeanah, 2014).
- Harlow's socially isolated monkeys should show intact sensory-motor function with devastated social cognition. **Confirmed:** isolated rhesus monkeys developed normal perceptual and motor skills but could not read social cues, predict others' behavior, or regulate their own emotional states (Harlow & Harlow, 1965).

Why this matters for the RSP model: This evidence constitutes a natural double dissociation. Neurological damage studies (Predictions 4 and 6) show that disrupting the neural substrates of schemas and self-modeling impairs consciousness even with social input available. Social deprivation studies (this prediction) show that removing social input impairs Level 3 consciousness even with intact neural substrates. Together, they demonstrate that both the neural architecture AND social calibration are necessary for full recursive social-predictive consciousness; neither alone is sufficient. This is the RSP model's strongest empirical differentiator from competing theories, which do not predict this specific pattern of selective Level 3 impairment under social deprivation.

X. Comparison with Alternative Views

A recent scoping review (Sattin et al., 2021) catalogued over thirty theoretical models of consciousness published between 2007 and 2017, spanning neural, cognitive, computational, quantum, and phenomenological approaches. In this section I situate RSP within this broader field, organized by theory family, and extend the comparison to include Cleeremans et al.'s (2020) SOMA framework, Bennett's (2025) Stack Theory, and Trukovich's (2025a, 2025b) Reaction to Reflection model, three post-review frameworks that address recursive self-modeling and consciousness from learning theory, computational theory, and theoretical biology respectively. For each comparison I identify points of alignment, points of divergence, and what RSP specifically adds. A synthesis table appears at the end.

A. Global Workspace Theories

Vs. Global Workspace Theory (GWT) and Global Neuronal Workspace (GNW)

The theory (Baars, 1988; Dehaene & Changeux, 2011): Consciousness arises when a single salient piece of information is selected, amplified through NMDA-mediated feedback connections causing "ignition," and globally broadcast across cortical networks via long-range pyramidal cell connections in layers II/III.

Neural correlates include prefrontal, cingulate, and parietal cortices. The transition from non-conscious to conscious processing is marked by a late, nonlinear amplification signature.

Alignment with RSP:

- Both frameworks treat consciousness as requiring integration across distributed neural networks, not localized to a single area.
- GNW's hierarchical feed-forward/feedback architecture is structurally compatible with RSP's three-level predictive hierarchy (Appendix A). RSP's precision-weighted prediction errors ascending the hierarchy and predictions descending it can be seen as a specific implementation of the workspace's broadcasting mechanism.
- Both predict that consciousness requires a threshold of activation. GNW's ignition threshold parallels RSP's minimum free energy reduction required for stable self-modeling (Appendix B, F_3).

Divergence from RSP:

- GNW explains *access consciousness* (how information becomes globally available for report and decision-making) but is largely silent on *phenomenal character*: why conscious experience feels like something. RSP addresses this directly through the recursive social prediction loop (Level 3), which generates the feeling of consciousness by turning the system's predictive machinery on itself.
- GNW treats all information equally once broadcast. RSP insists that the *content* matters: meaningful integration requires core affect (Level 1) structured through schemas (Level 2) and interpreted through concepts, not merely any information that wins the competition for broadcasting.
- GNW has no developmental account. RSP provides a mathematically grounded developmental progression (Section V) where Level 1 must stabilize before Level 2 can be commissioned, and Level 2 before Level 3, a cascade control constraint formalized in Appendix D.

What RSP adds: A specific mechanism for phenomenal character (recursive self-modeling via social prediction), a developmental ordering constrained by free energy minimization (Appendix B) and cascade control stability (Appendix D), and an explanation of *why* broadcasting feels like something: the content being broadcast includes a self-model that models itself being broadcast.

Vs. Semantic Pointer Architecture (SPA)

The theory (Eliasmith, 2013): Consciousness derives from interactive competition between semantic pointers, compressed neural representations of different natures (perceptive, motor, verbal). The winning pointer determines the qualitative character of experience.

Alignment: Both frameworks involve competitive selection among representations and emphasize binding across modalities. RSP's conceptual categorization (Level 3a) functions similarly to semantic pointer competition.

Divergence: SPA lacks an account of self-modeling or recursive prediction. It explains what content enters consciousness but not why consciousness feels like something. RSP's strange loop (Hofstadter, 1979), where the system predicts its own experience of predicting, goes beyond pointer competition.

What RSP adds: The recursive dimension. SPA's semantic pointers could serve as the representational format within RSP's three levels, but the strange loop that makes the system a subject rather than merely a processor is absent from SPA.

B. Higher-Order Representation Theories

Vs. Higher-Order Thought Theory (HOT)

The theory (Rosenthal, 1997; Lau & Rosenthal, 2011): A mental state is conscious when it is the object of a higher-order mental state directed at it. Consciousness requires meta-representation: representing that you are in a representational state. Neural correlates center on prefrontal cortex, with TMS studies showing PFC disruption affects awareness without impairing task performance.

Alignment with RSP:

- Both require self-directed representation for full consciousness. RSP's Level 3 recursive self-model is a form of higher-order representation.
- Both predict that prefrontal damage should selectively impair self-awareness while potentially preserving simpler forms of experience.

Divergence from RSP:

- HOT requires that the higher-order thought be *about* the first-order state, creating a two-level architecture. RSP proposes three levels with a recursive loop at Level 3, where the self-model models itself modeling, a richer structure than simple meta-representation.
- HOT implies that any system with higher-order thoughts is conscious, regardless of content. RSP insists that the content must be grounded in core affect (Level 1) and structured through schemas (Level 2). A system could have higher-order thoughts about arbitrary information without generating phenomenal consciousness.
- A critical problem for HOT (Gennaro, 2012): if PFC activity is necessary for all consciousness, then infants and most animals lack conscious experience entirely. RSP avoids this by granting simpler organisms Level 1 and Level 2 consciousness (core affect + schemas) without requiring Level 3 recursive self-modeling.
- HOT has no mathematical formalization. RSP's eight appendixes (A–H) provide free energy bounds, convergence proofs, stability analysis, and a formal theory of intelligence.

What RSP adds: A graded account of consciousness (Levels 1, 2, 3) that doesn't require meta-representation for basic phenomenal experience, a specific mechanism for how higher-order representation emerges (social prediction turned inward), and formal convergence guarantees (Appendix D, contraction mapping).

RSP's Level 3b is, in one sense, a higher-order representation: it is a representation of the system's own first-order states. But it differs from classical HOT (Rosenthal, 2005) in three respects. First, the higher-order state in RSP is specifically a social prediction (how others model me), not a generic higher-order thought. Second, the relationship between Level 3b and lower levels is constitutive rather than causal; the

higher-order state does not cause consciousness but partially constitutes it through integration with Levels 1-2. Third, RSP provides a principled account of WHY higher-order representations exist (social prediction demands required recursive self-modeling), whereas classical HOT treats the existence of higher-order thoughts as an unexplained given.

Vs. First-Order Representationalism (FOR)

The theory (Dretske, 1995; Tye, 2000): Consciousness is determined by perceptual representations formed in sensory regions. Specific consciousness (contents) depends on sensory structures; general consciousness depends on post-sensory structures in prefrontal and parietal cortex.

Alignment: RSP agrees that sensory representations play a foundational role. Level 1 core affect is essentially a first-order interoceptive representation, and RSP's Levels 1 and 2 are compatible with FOR's emphasis on sensory primacy.

Divergence: FOR claims first-order representations are sufficient for consciousness, making higher-order processing unnecessary. RSP argues that while Level 1 experience (core affect) may be "first-order" in character, the rich phenomenal consciousness humans experience requires the recursive social prediction loop of Level 3.

What RSP adds: An account of *how* first-order representations become conscious experience, through meaningful integration with schemas and recursive self-modeling, rather than simply asserting that they are.

Vs. Cross-Order Integration Theory (COI)

The theory (Kriegel, 2009, 2019): Conscious states arise from functional integration of a first-order representation with a higher-order representation of that first-order representation. The integration itself, not either representation alone, constitutes consciousness.

Alignment: COI's emphasis on *integration* across representational levels is highly compatible with RSP. RSP is fundamentally about meaningful integration, combining core affect (first-order) with schemas (structured representations) and recursive self-models (higher-order). COI's insistence that the binding itself matters, not just the components, mirrors RSP's central claim.

Divergence: COI proposes a two-level architecture (first-order + higher-order). RSP proposes three levels with an explicit recursive loop. COI lacks developmental, temporal, or mathematical specificity; it does not specify timescales, convergence properties, or how integration develops.

What RSP adds: RSP can be seen as a specific, formalized implementation of COI's integration principle, specifying *what* gets integrated (affect + schemas + self-model), *how* integration occurs (precision-weighted prediction error minimization, Appendix A), *when* integration stabilizes (timescale separation, Appendix D), and *why* it develops in a particular order (free energy constraints, Appendix B).

Vs. Attention Schema Theory (AST)

The theory (Graziano, 2013): Consciousness is a perceptual reconstruction of attentional state. The brain constructs an internal model of attention (the "attention schema") using the same social perception machinery (STS, TPJ, mPFC) that models other minds. When this machinery is directed at one's own attention, the result is the subjective sense of awareness.

Alignment with RSP:

- This is the closest competitor to RSP. Both theories ground consciousness in social cognition machinery turned inward. Both identify the same neural substrates (TPJ, STS, mPFC). Both predict that social cognition deficits should correlate with altered self-awareness.
- RSP explicitly incorporates an attention schema as one of its five core schemas at Level 2 (Section III), acknowledging Graziano's contribution.

Divergence from RSP:

- AST treats consciousness as a *model of attention*, a specific, relatively narrow perceptual reconstruction. RSP treats consciousness as a recursive social prediction loop that encompasses not just attention but the full self-model including body schema, spatial schema, affective-homeostatic schema, and self-schema, all integrated through prediction error minimization.
- AST does not emphasize recursion. The attention schema models attention, but it does not recursively model itself modeling attention. RSP's strange loop, where the self-model predicts its own future states including its predictions about its predictions, is what generates the distinctive quality of conscious experience.
- AST lacks formal mathematical treatment. RSP's contraction mapping proof (Appendix D) shows that the recursive loop converges to a stable fixed point $\mathcal{M}^*$, avoiding the infinite regress that plagues many self-referential theories. AST provides no such guarantee.
- AST does not account for core affect or the interoceptive foundation of consciousness. RSP grounds everything in Level 1 valence and arousal, explaining why consciousness always has an affective tone.
- AST has no developmental account. RSP's Level 1 $\rightarrow$ 2 $\rightarrow$ 3 developmental sequence, grounded in the mathematical necessity of bottom-up construction (Appendices B, C, D), explains why children develop bodily awareness before attentional awareness before recursive self-awareness.

What RSP adds: AST identifies one piece of the puzzle, the attention schema, and correctly notes its social origins. RSP incorporates this insight within a larger architecture that includes four additional schemas, an interoceptive foundation, recursive self-modeling with convergence guarantees, developmental ordering, and four mathematical formalizations. AST explains why we think we are conscious; RSP explains why consciousness feels like something.

Vs. Self-Organizing Metarepresentational Account (SOMA; Cleeremans et al., 2020)

The theory (Cleeremans et al., 2020): SOMA proposes that consciousness is something the brain *learns* to do. Through continuous representational redescription (RR), the brain unconsciously learns to re-represent its own first-order activity, generating metarepresentations that qualify those states as

conscious. The framework integrates elements of GWT (global broadcasting), HOT (higher-order monitoring), social theories (the self-other loop), and predictive processing (prediction-driven learning). SOMA identifies three entangled loops: an *inner loop* through which the brain redescribes its own activity, a *perception-action loop* through which agents learn about the consequences of their actions, and a *self-other loop* through which they learn about other agents. Experiences occur only in systems that have "learned to know they possess certain first-order states and that have learned to care more about certain states than about others." Consciousness, on this account, is "the brain's (unconscious, embodied, enactive, non-conceptual) theory about itself."

Alignment with RSP:

- Both frameworks identify social cognition as playing a constitutive role in consciousness, not merely a developmental one. SOMA's self-other loop and RSP's Level 3b both ground consciousness in mechanisms originally evolved for modeling other agents.
- Both propose hierarchical, multi-loop architectures with an inner monitoring system augmented by external interaction loops. SOMA's three loops (inner, perception-action, self-other) roughly parallel RSP's three levels, though the inner loop's metacognitive monitoring is closer to a precursor of Level 3b than to the interoceptive regulation of Levels 1-2.
- Both draw on predictive processing. SOMA's prediction-driven learning mechanisms and RSP's hierarchical predictive coding (Appendix A) share the core insight that the brain is a prediction machine that generates consciousness through self-directed prediction.
- Both deny that consciousness is an intrinsic property of individual neural states, though they differ sharply on what replaces that view: SOMA emphasizes learning, RSP emphasizes recursive architecture.

Divergence from RSP:

- **Epistemic vs. ontological claim.** SOMA is fundamentally an epistemic theory: consciousness arises when the brain *learns to know* its own states. RSP makes an ontological claim: phenomenology IS the recursive self-modeling process, not the brain's knowledge about that process. This is a first-order philosophical difference. SOMA explains how the brain becomes *aware* of its states; RSP claims the self-modeling process *constitutes* experience. A system could, in principle, satisfy SOMA's conditions (having learned metarepresentations of its first-order states) without recursive social prediction; RSP denies such a system would be phenomenally conscious.
- **Affect as architectural foundation vs. selection criterion.** SOMA treats motivational valence as a criterion for which states become conscious ("learned to care more about certain states"), making it a selector among learned metarepresentations. RSP makes core affect (Level 1) architecturally foundational: the entire predictive hierarchy is grounded in valence and arousal, and the strange loop is always affectively toned. The difference is structural, not merely terminological. RSP predicts that disrupting Level 1 affect will fundamentally alter or eliminate conscious experience; SOMA, where affect serves as a selection filter rather than the architectural base, makes no such prediction.

- **Mathematical formalization.** SOMA is entirely conceptual, containing no equations, no formal dynamics, and no convergence proofs. RSP provides five core mathematical formalisms (Appendixes A-E) with proven cross-formalism equivalences, a formal theory of intelligence (Appendix H), and an algorithmic specification (§E.9). This means RSP generates precise, falsifiable quantitative predictions that SOMA cannot.
- **Developmental specificity.** SOMA inherits general developmental predictions from Karmiloff-Smith's representational redescription framework, but does not specify when metarepresentational consciousness emerges or in what order its three loops come online. RSP specifies the construction order (Level 1 by 6 months, Level 2 by 24 months, Level 3 emerging at 4-5 years, full recursion by 7+) and grounds it in mathematical necessity (cascade control stability, free energy reduction ordering, reward-signal prerequisites).
- **Clinical predictions.** SOMA does not map clinical dissociations onto architectural failures. RSP provides a systematic clinical prediction matrix (Appendix G) mapping specific conditions (Cotard's, depersonalization, Capgras, DID) onto level-specific disruptions with testable neuroimaging signatures.
- **Phylogenetic predictions.** SOMA's learning criterion is species-permissive: any system with sufficient representational redescription capacity could develop consciousness, potentially including cephalopods or even sufficiently plastic artificial systems. RSP's social-prediction requirement is more restrictive: Level 3b requires recursive social cognition, predicting that solitary species with rich learning (octopuses) lack phenomenal self-awareness despite sophisticated Level 1-2 capacity (§V). This is an empirically adjudicable difference.
- **The hard problem.** SOMA does not claim to dissolve or even address the hard problem directly. RSP defends a strong dissolution: phenomenology is constitutive of recursive self-modeling, not a separate property that could be absent from it (§IV).

What RSP adds: The constitutive identity claim (phenomenology IS recursive self-modeling, not the brain's learned knowledge about itself), an interoceptive foundation that makes affect architecturally primary, five mathematical formalisms, developmental timelines grounded in mathematical necessity, a clinical prediction matrix, and a dissolution of the hard problem. SOMA explains how metarepresentation is learned; RSP explains why learning it should feel like anything at all.

What SOMA adds to RSP: The emphasis on learning and plasticity as mechanisms through which metarepresentational capacity develops. RSP's developmental account describes the *order* of level construction but says less about the *learning mechanisms* that build each level. SOMA's representational redescription framework could provide a computational implementation for how RSP's schemas and self-models are acquired through experience.

Overall relationship: Extends (with a caveat). RSP incorporates SOMA's structural insights (multi-loop architecture, social-origin of self-modeling, prediction-driven learning) while replacing its epistemic core with an ontological constitutive identity claim. This is "extends" rather than "complements" because SOMA's architectural elements are preserved within RSP's larger framework, even though the philosophical interpretation shifts from "the brain learns to know its states" to "the self-modeling process constitutes experience." RSP adds the mathematical formalism, interoceptive foundation, developmental

specificity, clinical predictions, and hard-problem dissolution that SOMA lacks. SOMA’s core insight (consciousness involves learned metarepresentation) is compatible with RSP but insufficient: it does not specify *what* is learned, *why* the learning produces phenomenality rather than mere monitoring, or *how* to test the theory quantitatively.

C. Information Integration Theories

Vs. Integrated Information Theory (IIT)

The theory (Tononi, 2008; Oizumi et al., 2014): Consciousness is identical to integrated information (Φ). Any system with sufficient Φ is conscious to a degree. Consciousness has both quantity (Φ_{max}) and quality (the shape of informational relationships in qualia space). IIT derives from five phenomenological axioms (existence, composition, information, integration, exclusion) mapped to five ontological postulates. Key neural correlate: the cortico-thalamic system (high integration), not the cerebellum (high neuron count but low integration).

Alignment with RSP:

- Both theories take mathematical formalization seriously. IIT’s Φ calculus and RSP’s five core mathematical appendixes (A–E) represent the most formally developed approaches in the field.
- Both insist that consciousness admits of degrees, not an all-or-nothing phenomenon.
- Both identify the cortico-thalamic system as critical.
- Both predict that integration across neural populations, not mere neural activity, determines consciousness.

Divergence from RSP:

- IIT is *architecture-neutral*: any system with sufficient Φ is conscious, regardless of its physical organization or what information is integrated. (IIT does have content via its qualia space, where the shape of integrated information determines *what* a system experiences, but it is agnostic about *which architectures* can be conscious.) RSP is *architecture-specific*: consciousness requires particular content (core affect, schemas, and recursive self-models) integrated in a particular three-level architecture. A thermostat with nonzero Φ is conscious under IIT; under RSP it is not.
- IIT’s Φ is computationally intractable for biological systems (exponential in system size). RSP’s free energy formulation (Appendix B) uses gradient descent on a tractable variational bound, making it more amenable to empirical measurement.
- IIT starts from phenomenological axioms and works backward to physical substrates. RSP starts from neuroscience (interoception, predictive coding, social cognition) and works forward to explain phenomenology. The two approaches arrive at different conclusions precisely because they differ in starting point.
- IIT has no developmental account. RSP explains why consciousness develops in a specific order, grounded in the mathematical necessity of stabilizing inner control loops before outer ones (Appendix D).

- IIT predicts that cerebellum lesions should not affect consciousness (low integration despite high neuron count). RSP agrees but for a different reason: the cerebellum lacks the recursive social prediction circuitry of Level 3, not merely because its Φ is low.

What RSP adds: A mechanism for *why* integration generates consciousness (the recursive social prediction loop), a developmental progression, a content-specific account of phenomenal character, and a tractable mathematical framework. IIT tells us *that* integration matters; RSP tells us *which* integration matters and *why*. Notably, the adversarial collaboration between GNW and IIT (Melloni et al., 2023) found evidence favoring a posterior cortical "hot zone" over GNW's frontoparietal ignition signature. RSP's multi-level architecture may bridge this debate: Level 1–2 substrates (subcortical, posterior cortical) align with IIT's posterior localization, while Level 3 (prefrontal, TPJ, mPFC) aligns with GNW's frontal emphasis — suggesting both are partially correct about different levels of the architecture.

IIT's exclusion postulate, which holds that only the maximum of integrated information (the "complex") is conscious, directly challenges RSP's multi-level architecture, since RSP claims consciousness spans three levels with different timescales. RSP's response: the exclusion postulate is an axiom, not an empirical finding. The clinical evidence (Prediction 6) suggests that consciousness can be partially disrupted at specific levels, producing graded rather than all-or-nothing changes in phenomenology. If the exclusion postulate were correct, Level 3b damage should either eliminate all consciousness or leave it entirely intact. That depersonalization, Cotard's delusion, and thought insertion produce graded intermediate states suggests otherwise.

Vs. Adaptive Resonance Theory (ART)

The theory (Grossberg, 2021): Consciousness arises from resonant states produced by matching between bottom-up sensory inputs and top-down expectations. Only resonant states that attract attention and have been learned constitute conscious experiences. Key predictions: attentional matching associated with gamma oscillations; mismatch processing with beta oscillations.

Alignment with RSP:

- Both are fundamentally predictive processing frameworks: bottom-up signals meet top-down predictions, and the match/mismatch determines what enters consciousness. RSP's prediction error minimization (Appendix A) and ART's resonance are computationally related.
- Both address the stability-plasticity dilemma (how the system learns new patterns without forgetting old ones). RSP's timescale separation ($\tau_1 \ll \tau_2 \ll \tau_3$, Appendix D) serves a similar function to ART's resonance-mediated learning.
- Both incorporate emotional processing as integral to consciousness, not peripheral.

Divergence: ART focuses on the perceptual matching process itself. RSP asks what the content of that matching is (core affect + schemas + self-model) and adds the recursive self-referential dimension. ART explains *how* information becomes conscious; RSP adds *why* it feels like something by including the strange loop.

What RSP adds: The social-predictive dimension. ART's resonance could serve as the neural mechanism implementing RSP's prediction error minimization at Levels 1 and 2, but ART lacks the recursive self-modeling of Level 3 that generates subjective experience.

D. Hierarchical and Developmental Theories

Vs. Consciousness State Space Model (CSS)

The theory (referenced as Sattin et al.'s framework): Consciousness is organized along three continuous dimensions (time, awareness, and emotion) with two interrelated categories: core consciousness (here-and-now, agency, ownership) and extended consciousness (episodic memory, prospective thought, verbal reflection). The emotion dimension subdivides into valence (extended) and arousal (core).

Alignment with RSP:

- Both propose multi-dimensional frameworks with affect (valence/arousal) as a foundational dimension.
- CSS's core-vs-extended consciousness distinction maps roughly onto RSP's Level 1-2 (core) vs. Level 3 (extended through recursive social prediction).
- Both identify the default mode network with extended/recursive consciousness.

Divergence: CSS describes the *space* of conscious states but does not specify the *mechanism* that generates them. RSP provides the mechanism: hierarchical predictive coding with recursive self-modeling. CSS's three dimensions (time, awareness, emotion) are descriptive coordinates; RSP's three levels (core affect, schemas, recursive social prediction) are causal mechanisms with formal dynamics.

What RSP adds: A causal, mechanistic account of how the CSS dimensions arise. Emotion from Level 1, awareness from Level 2 schemas, and temporal extension from Level 3 predictive self-modeling. RSP also provides mathematical formalization (Appendix B) and stability analysis (Appendix D).

Vs. Psychological Theory of Consciousness (PToC)

The theory (Shanon, 2008): A tripartite hierarchy comprising Cons1 (sentience/sensed being), Cons2 (mental awareness/subjective experience), and Cons3 (meta-mentation/reflection). Includes non-ordinary states (Cons4: hallucinations; Cons5: mystical/transcendent).

Alignment: Striking structural parallel with RSP. Cons1 $\approx$ Level 1 (core affect/sentience), Cons2 $\approx$ Level 2 (structured experience through schemas), Cons3 $\approx$ Level 3 (recursive self-reflection). Both predict that these levels develop hierarchically.

Divergence: PToC is purely descriptive and psychological. It lacks neurobiological grounding, mathematical formalization, or a specific mechanism for how the levels relate. RSP provides all three: neural substrates for each level, four mathematical formalisms, and the recursive social prediction mechanism.

What RSP adds: RSP can be understood as a neurobiological and mathematical realization of PToC's intuitive hierarchy. Where PToC describes, RSP explains.

E. Neural Architecture Theories

Vs. The Centrencephalic Proposal (CP)

The theory (Merker, 2007, developing Penfield and Jasper): Consciousness is determined by subcortical brainstem activity involving real-time simulation with target selection, action selection, and motivational ranking in an ego-centric frame. The superior colliculus is the key structure, along with the periaqueductal gray (PAG), due to their bidirectional cortical connections and multisensory integration.

Alignment: RSP agrees that subcortical structures are essential. Level 1 core affect depends on brainstem, hypothalamus, amygdala, and anterior insula, precisely the structures CP emphasizes. Both predict that brainstem lesions should fundamentally alter consciousness.

Divergence: CP downplays cortical contributions. RSP argues that while Level 1 (subcortical core affect) is foundational, the rich phenomenal consciousness humans experience requires cortical Level 2 schemas and Level 3 recursive social prediction. CP provides a necessary but insufficient account.

What RSP adds: RSP incorporates CP's subcortical foundation as Level 1 while explaining what cortical processing adds: structured experience (Level 2) and recursive self-awareness (Level 3). The cascade control formalization (Appendix D) specifies exactly how subcortical and cortical systems interact across timescales.

Vs. Recurrent Processing Theory (RPT)

The theory (Lamme, 2006): Consciousness requires recurrent (feedback) processing between cortical areas. Feed-forward processing alone is unconscious; recurrent processing generates phenomenal consciousness. This predicts that interrupting recurrent loops (e.g., through backward masking) should eliminate consciousness of a stimulus.

Alignment: RSP is deeply compatible with RPT. The entire RSP architecture is built on recurrent processing: prediction errors ascending, predictions descending, and the Level 3 strange loop as the ultimate recurrent circuit. RSP's predictive coding formalization (Appendix A) is an explicit specification of what recurrent processing computes.

Divergence: RPT identifies recurrent processing as necessary and sufficient for consciousness but does not specify *what content* the recurrent loops carry. RSP specifies that the relevant recurrent loops carry core affect, schema predictions, and recursive self-models, not just any recurrent visual processing.

What RSP adds: Content specificity. RPT tells us the computational architecture (recurrent loops); RSP tells us what those loops compute (meaningful integration through social prediction).

F. Quantum Consciousness Theories

Vs. Orchestrated Objective Reduction (Orch OR)

The theory (Penrose, 1989; Penrose & Hameroff, 1996): Consciousness consists of discrete moments of objective reduction (OR) of quantum states in brain microtubules. Quantum coherent superposition

evolves during integration phases until a threshold at $\tau \approx \hbar/E_G$ produces a conscious moment. Synaptic inputs "orchestrate" these quantum computations.

Alignment: Both theories propose specific timescales for conscious processing. Orch OR's τ for quantum collapse and RSP's $\tau_1 \approx 100ms$ for core affect processing are in comparable ranges. Both take mathematical formalization seriously.

Divergence: Fundamental disagreement on level of explanation. RSP operates at the neural circuit and systems level using classical computation (predictive coding, control theory, RL). Orch OR requires non-computational quantum processes at the sub-neuronal level. RSP argues that consciousness can be fully explained by classical dynamical systems; the contraction mapping convergence proof (Appendix D) demonstrates that the strange loop stabilizes without invoking quantum mechanics. The empirical challenge for Orch OR remains severe: quantum decoherence times in warm, wet brain tissue ($\sim 10^{-13}$ seconds for neural superpositions; Tegmark, 2000) appear far too short to sustain the coherence required, though Hameroff and Penrose have argued that microtubule geometry could provide partial shielding.

What RSP adds: A fully classical, biologically plausible account that does not require speculative quantum coherence in microtubules. RSP's mathematical apparatus (variational free energy, cascade control, contraction mapping) achieves formal rigor comparable to quantum mechanics but at a level of description with clear neural correlates.

Vs. Quantum Consciousness Theories Generally

The scoping review identified six quantum theories of consciousness. While they vary in mechanism (wave function collapse, quantum holography, single-particle panpsychism, no-go theorems), they share a common claim: classical neural computation is insufficient for consciousness.

RSP's response: RSP demonstrates that classical computation is sufficient when organized as hierarchical predictive coding with a recursive self-referential loop. The five core mathematical appendixes (A–E) show that the properties often attributed to quantum processes (non-decomposability in Appendix B's integrated free energy, emergent dynamics in Appendix D's strange loop convergence, and subjective privacy in the intrinsically first-personal self-model) arise naturally from classical dynamical systems theory. The parsimony principle favors RSP: quantum consciousness theories require mechanisms with no established empirical support in neural tissue, while RSP uses well-established neuroscience.

G. Emotion and Affect Theories

Vs. Biological Universalism (Basic Emotions Theory)

The theory (Ekman, 1992; Panksepp, 1998): A small number of innate, universal emotions exist with dedicated neural circuits and cross-culturally recognizable expressions. Note: Ekman and Panksepp represent distinct theoretical commitments grouped here for comparison purposes. Ekman (1992) posits discrete universal facial expressions mapped to emotion categories — a surface-level, expression-based account. Panksepp (1998) posits seven primary subcortical emotional systems (SEEKING, RAGE, FEAR,

LUST, CARE, PANIC/GRIEF, PLAY) grounded in affective neuroscience. Panksepp's subcortical systems are architecturally closer to RSP's Level 1 core affect than to Ekman's discrete-expression model.

Alignment: RSP agrees that core affect systems (approach/avoidance, pleasure/distress) are universal and subcortical. This is precisely Level 1. Panksepp's subcortical emotional circuits in particular map well onto the Level 1 substrate.

Divergence: RSP treats these subcortical systems as generating dimensional affect, not discrete emotions. Specific emotions require conceptual categorization through Level 2 schemas and Level 3 cultural learning. RSP diverges from Ekman more sharply than from Panksepp: Ekman's discrete-expression model conflicts with RSP's constructionist account of emotion categories, while Panksepp's subcortical systems are largely compatible with RSP's Level 1.

Evidence: Emotion recognition is not as universal as Ekman claimed (Jack et al., 2012); supposed basic emotions show high within-category variation (Barrett, 2006).

Vs. Strong Constructionism and Embodied Appraisal

The theories: Two distinct positions are relevant here. Barrett's (2017) *theory of constructed emotion* holds that discrete emotions are constructed from the combination of core affect (a biological primitive) with learned conceptual categories; she does not deny innate affective systems but denies innate discrete emotions. Prinz's (2004) *embodied appraisal theory* holds that emotions are perceptions of bodily changes (a neo-Jamesian view), emphasizing somatic markers rather than cognitive construction.

Alignment with Barrett: RSP's three-level architecture is closely aligned with Barrett's framework. RSP's Level 1 (core affect) corresponds to Barrett's dimensional affect; RSP's Level 3a (conceptual categorization) corresponds to Barrett's constructed emotion categories. The key agreement: concepts partially constitute emotional experience, not merely label it.

Alignment with Prinz: RSP's emphasis on interoception and bodily states at Level 1 resonates with Prinz's embodied appraisal account.

Divergence: RSP goes beyond both by adding (1) the intermediate Level 2 (core schemas) as structural organizers, (2) the recursive social prediction mechanism at Level 3b, and (3) the four mathematical formalizations. Barrett's theory lacks the recursive self-modeling component; Prinz's theory lacks the constructionist dimension.

Evidence: Core affect persists without cortical processing (Merker, 2007); subcortical structures are necessary for emotion (Damasio, 1994); but conceptual categories shape the specificity of emotional experience (Barrett et al., 2001).

RSP integrates the strongest elements of both positions: biological primitives (core affect, Level 1) organized through schemas (Level 2) and elaborated through constructed categories and recursive self-modeling (Level 3).

It should be stated plainly: Level 3a of the RSP architecture is Barrett's theory of constructed emotion. The conceptual categorization process, in which core affect is transformed into discrete emotional experience

through culturally learned concepts, is Barrett’s central insight, and RSP inherits it directly. What RSP adds is the Level 2 intermediate layer (five schemas that organize experience between raw affect and conceptual categorization) and the Level 3b recursive loop (social prediction redirected at the self). RSP should be understood as extending Barrett’s framework, not as an independent development.

H. Phenomenological Theories

Vs. Gurwitsch’s Theory (GT)

The theory (Gurwitsch, 1964; extended by Yoshimi & Vinson, 2015): Consciousness is organized into three domains: theme (focus of attention), thematic field (materially related data), and marginal consciousness (irrelevant co-present data). Passage between domains follows intrinsic content relationships.

Alignment: GT’s theme-field-margin structure can be mapped onto RSP’s precision weighting. The theme corresponds to the highest-precision prediction error signal; the thematic field to moderate-precision signals in the same schema domain; the margin to low-precision background predictions (Appendix A).

Divergence: GT is phenomenological: it describes the structure of experience without specifying the mechanism that generates it. RSP provides that mechanism through hierarchical precision-weighted prediction error minimization.

What RSP adds: A computational explanation for why experience has the theme-field-margin structure GT describes. The precision weighting formalized in Appendix A naturally generates focal, peripheral, and marginal awareness as a consequence of Bayesian inference.

I. Computational Intelligence Theories

Vs. Stack Theory of Intelligence and Consciousness (Bennett, 2025)

The theory (Bennett, 2025): Michael Timothy Bennett’s doctoral thesis at the Australian National University proposes a thorough theory linking intelligence to consciousness through a formalism called *Pancomputational Enactivism* within *Stack Theory*. The core argument proceeds in stages. First, every physical system is an abstraction layer: software is a state of hardware, hardware is a state of physics, and so on in an infinite stack. Second, intelligence is formalized as *w-maxing* (choosing the weakest correct policy, the one that applies to the broadest range of situations), which Bennett proves is strictly optimal for generalization, outperforming Ockham’s Razor (simp-maxing) by 110–500% in experiments. Third, systems learn pre-linguistic classifiers called *causal-identities* for objects that cause valence, which explains how a contentless environment gets divided into objects and properties. Fourth, consciousness arises through orders of self: a *1st-order-self* (reaffERENCE, classifying one’s own interventions, present even in flies), a *2nd-order-self* ("my prediction of your prediction of me," needed for communication and theory of mind), and a *3rd-order-self* (predicting one’s own 2nd-order selves, enabling complex social reasoning). Bennett proposes that phenomenal consciousness begins at the 1st-order-self, while access consciousness requires a 2nd-order-self.

Two further results are noteworthy. Bennett introduces the *Psychophysical Principle of Causality*: qualia are "tapestries of valence," vast ensembles of cells simultaneously attracted or repelled, whose multidimensional pattern constitutes the qualitative character of experience (what makes thirst feel different from hunger). And he identifies an open problem he calls *The Temporal Gap*: whether a conscious state must be realized synchronously at a point in time (Option 1, implying current sequential computers cannot be conscious) or can be "smeared across time" (Option 2, permitting software consciousness). Bennett favors Option 1.

Alignment with RSP:

- Both frameworks arrive at a hierarchical architecture with three orders of self-representation. Bennett's 1st/2nd/3rd-order selves map loosely onto RSP's three levels, though the boundaries differ significantly (see Divergence below). Bennett's 1st-order-self (reafference) overlaps with, though is broader than, RSP's Level 2 self-schema — reafference circuits are distributed across brainstem, cerebellum, and cortex, while RSP's self-schema targets cortical regions (mirror system, TPJ, precuneus); his 2nd-order-self ("my prediction of your prediction of me") maps most directly onto RSP's Level 3b recursive social prediction; his 3rd-order-self (predicting one's own 2nd-order selves) corresponds to deeper recursion within Level 3b.
- Both make valence foundational. Bennett's "tapestries of valence," in which representation and value judgement are inseparable, aligns directly with RSP's core affect (Level 1), where $\mathbf{a}(t) = (v(t), \alpha(t))$ grounds the entire predictive hierarchy (Appendix A, §A.3.1). Neither framework permits platonic, value-neutral representations.
- Both make philosophical zombies impossible. Bennett argues that human-level intelligence requires 1st and 2nd-order selves, which entail consciousness. RSP argues that the recursive self-modeling constituting consciousness is the same process constituting social-adaptive intelligence: the master functional $\mathcal{J}$ whose minimization realizes both cannot be optimized for one without realizing the other (see Appendix H for the formal argument).
- Both reject the separation of intelligence from goals. Bennett explicitly refutes the orthogonality thesis (Bostrom, 2014); RSP's constitutive identity proposition (Appendix H) constrains the thesis for RSP-type systems, since Level 1 valence is architecturally inseparable from the predictive hierarchy that constitutes intelligence.
- Both identify causal learning as essential. Bennett derives causal-identities from the need to discriminate interventions from observations; RSP formalizes this through the active inference mechanism (Appendix B, §B.4.3), which plays an analogous role to Pearl's do-operator by distinguishing self-generated from externally caused sensory changes.
- Both conclude that current AI lacks the architecture for consciousness (Bennett: no delegated *polycomputation*, his term for distributed, parallel, hierarchically organized computation where the same matter participates in multiple computations at different scales simultaneously; RSP: no embodied hierarchical predictive coding with recursive self-modeling).

Divergence from RSP:

- **Where phenomenal consciousness begins.** Bennett holds that phenomenal consciousness begins at the 1st-order-self — a fly that distinguishes its own movements from external events already has "something it is like." RSP holds that full phenomenal consciousness requires the recursive strange loop at Level 3b. RSP grants that organisms with only Levels 1–2 have valenced, structured experience (not nothing), but the rich phenomenal consciousness Nagel asks about requires the recursive self-model. This is an empirically adjudicable difference: RSP predicts flies lack phenomenal self-awareness; Bennett predicts they possess it.
- **Reductive vs. non-reductive physicalism.** Bennett is an explicit reductive physicalist in the sense that he aims to "break qualia down into something more basic" (tapestries of valence at lower abstraction layers), though his stack theory retains the causal efficacy of higher-level descriptions, placing his view closer to non-reductive positions than the label suggests. RSP is non-reductive: consciousness is constituted by physical processes organized in a specific way, but the organizational description is causally indispensable. RSP dissolves the hard problem through constitutive identity rather than reducing it.
- **Social prediction as constitutive vs. consequential.** RSP's central claim is that social prediction is *constitutive* of consciousness — the social prediction machinery turned inward IS the mechanism. Bennett treats social cognition (2nd-order selves) as a *consequence* of sufficient scale and incentive, not the constitutive core. For Bennett, consciousness begins before social cognition; for RSP, full consciousness requires it.
- **Formalism.** Bennett's mathematics is set-theoretic (policies as sets, extensions as cardinalities, proofs about optimal proxies). RSP's is dynamical-systems-based (differential equations, free energy functionals, contraction mappings). The two formalisms are complementary but not directly translatable, though Appendix H (§H.3.6) develops the generalization structure that parallels Bennett's weakness hierarchy.
- **Content specificity.** Bennett's w-maxing is content-neutral: it applies to any task in any vocabulary. RSP specifies exactly what an intelligent-and-conscious system must represent: core affect, five schemas, conceptual categories, and a recursive self-model. This yields falsifiable content-specific predictions that Bennett's framework does not.

What RSP adds: RSP provides what Bennett's framework lacks: neurobiological grounding (specific neural substrates for each level), a formal developmental progression with age-specific timelines (§V), content-specific predictions (§IX), and a dissolution of the hard problem rather than a reduction. RSP also offers a potential answer to Bennett's Temporal Gap: the timescale separation requirement ($\tau_{l+1}/\tau_l \geq 5$; Appendix D) specifies that consciousness requires real-time processing within level-specific temporal windows, providing independent motivation for Bennett's Option 1, though timescale separation (temporal hierarchy) is not identical to the synchronous instantiation Bennett's Option 1 requires.

What Bennett adds to RSP: Bennett provides an independent computational-theoretic argument for the same hierarchical architecture, a formal proof that weakness (generalization breadth) is the unique optimal learning criterion, the *polycomputation* argument for why biological substrate supports consciousness better than sequential computation, and The Temporal Gap as a genuinely novel

philosophical problem that RSP should engage with. His proof that intelligence requires consciousness, reached from AI theory rather than neuroscience, provides an independent argument for a conclusion RSP reaches from different premises (Appendix H, §H.4). Both draw on predictive coding and social cognition literatures, so the independence is partial rather than complete, but the formalisms and starting points are distinct. The conjecture that RSP's free energy minimization may be the biological realization of Bennett's w-maxing is developed in Appendix H (§H.3.6).

Overall relationship: Complements. Bennett's Stack Theory and RSP arrive at convergent conclusions from independent starting points — computational learning theory and hierarchical predictive coding respectively. The two frameworks are mutually reinforcing: Bennett provides the optimality proof RSP lacks; RSP provides the neurobiological specificity Bennett lacks. Together, they suggest that the hierarchical, valence-grounded, recursively self-modeling architecture is not merely one way to build a conscious mind but provably optimal for adaptive intelligence in social environments.

J. Recursive/Evolutionary Theories

Vs. Reaction to Reflection Model (R2R; Trukovich, 2025a, 2025b)

The theory: Joseph Trukovich's Reaction to Reflection (R2R) model, developed across two papers in *BioSystems* (2025a on the evolution of cognition and complexity; 2025b on the structure of intelligence, with thematic overlap with the computational intelligence theories in §I), proposes that intelligence and consciousness emerge through increasing *recursive depth*, the capacity for self-referential processing grounded in biological systems with authentic thermodynamic constraints. R2R identifies four evolutionary transitions: *reaction* (chemical recursion in metabolic networks), *temporogenesis* (anticipatory prediction through internal temporal models), *symbiogenesis* (cooperative integration; Trukovich uses this term more broadly than Margulis's endosymbiotic origin of organelles, referring to cooperative integration across organisms generally), and *cognogenesis* (explicit self-referential modeling). The central distinction is between *implicit recursion*, where systems engage recursive processes without representing them, and *explicit recursion*, where recursion itself becomes a manipulable cognitive construct. Consciousness emerges at this transition. Trukovich introduces *Value Saturation*: phenomenal consciousness IS explicit recursive self-modeling saturated by homeostatic significance under *perspectival entrapment* — the inescapable constraint of occupying a particular biological viewpoint (a concept with antecedents in Metzinger's (2003) *perspectivalness* as a property of phenomenal self-models and Zahavi's (2005) minimal self, though the "entrapment" framing, emphasizing constraint rather than ground, is Trukovich's own gloss on these themes).

Alignment with RSP:

- Both frameworks make an identity claim: consciousness IS recursive self-modeling, not a product of it. RSP's constitutive identity (§IV) and R2R's Value Saturation converge on the same philosophical move — dissolving the explanatory gap by identifying phenomenal experience with the recursive process itself, grounded in biological affect.

- Both ground consciousness in homeostatic significance. RSP's core affect (Level 1) and R2R's "authentic thermodynamic constraints" (by which Trukovich appears to mean non-equilibrium metabolic processes that create genuine survival stakes, though the term is not formally defined) play the same architectural role: recursion without biological grounding lacks the evaluative significance that constitutes experience.
- Both reject substrate independence. RSP requires the specific three-level recursive architecture; R2R requires biological thermodynamic grounding. Neither permits consciousness in arbitrary computational substrates.
- Both trace an evolutionary progression from simple to complex recursion, with consciousness emerging at the self-referential stage. R2R's four transitions (reaction → temporogenesis → symbiogenesis → cognogenesis) roughly parallel RSP's three levels, though the mapping is not exact: R2R's symbiogenesis (cooperative integration) has no direct RSP analogue, while RSP's social prediction mechanism is absent from R2R.
- The convergence is notable because researchers working from methodologically distinct traditions (predictive coding/neuroscience vs. theoretical biology/complex systems) arrived at structurally similar identity claims. The independence is partial rather than complete, since both traditions draw on recursive self-modeling concepts traceable to Hofstadter (1979) and Metzinger (2003), but the formalisms and starting points differ enough that the convergence provides theoretical consilience beyond what either framework achieves alone.

Divergence from RSP:

- **Social prediction as constitutive.** RSP's central claim is that *social* prediction turned inward constitutes the strange loop. R2R derives consciousness from recursive depth generally — any system with sufficient self-referential complexity under biological constraints. RSP makes a stronger, more specific, and therefore more falsifiable claim.
- **Perspectival entrapment vs. architectural functionalism.** R2R argues that being trapped within a biological perspective is constitutive of consciousness. RSP argues that the specific three-level architecture is constitutive. These are different commitments: RSP could in principle be realized in a non-biological substrate with the right architecture; R2R denies this.
- **Mathematical formalism.** RSP provides five core mathematical appendixes (A–E: predictive coding, free energy, RL, control theory, unified framework) with explicit dynamics, convergence proofs, and falsification conditions, plus a formal theory of intelligence (Appendix H). R2R is primarily conceptual, drawing on evidence from microbial decision-making, cross-kingdom signaling, and neural networks without a unified dynamical-systems formalism.
- **Content specificity.** RSP specifies exactly what must be represented (core affect, five schemas, recursive self-model) and derives falsifiable predictions. R2R characterizes consciousness through recursive depth without specifying the content of the recursion.
- **Developmental predictions.** RSP provides age-specific developmental timelines tied to level-specific emergence (§V). R2R's four evolutionary transitions operate on phylogenetic rather than ontogenetic timescales.

What RSP adds: Neurobiological specificity (named substrates for each level), content-specific predictions (§IX), mathematical formalism (Appendixes A–E), developmental timelines (§V), and the social prediction mechanism that explains *why* the strange loop emerges in social species.

What R2R adds to RSP: The concept of Value Saturation, the claim that recursive self-modeling must be *saturated* by homeostatic significance to constitute consciousness, resonates with RSP's insistence that the strange loop is always grounded in Level 1 affect, and may offer a more explicit formulation of why affect-grounding is necessary rather than contingent. The related concept of perspectival entrapment offers a candidate explanation for why recursive self-modeling feels like something from the inside. The symbiogenesis stage raises a question RSP should address: whether cooperative (non-recursive) integration across organisms constitutes a distinct evolutionary achievement or reduces to RSP's Level 2 schema elaboration. And R2R's grounding in theoretical biology (metabolic networks, cross-kingdom signaling) extends the evolutionary story below the bilaterian valence systems where RSP's Level 1 narrative begins (§V traces core affect to ~600 Mya), into pre-neural and single-celled recursive processes.

A note on dissolution: It is worth asking whether R2R achieves the same strong dissolution RSP claims. RSP's constitutive identity thesis closes the explanatory gap by identifying phenomenology with the recursive process itself. R2R's perspectival entrapment could be read as an additional explanatory primitive. If the *feel* of consciousness requires being trapped in a biological perspective, the question "why does perspectival entrapment feel like anything?" threatens to reopen the gap. RSP can absorb perspectival entrapment as a consequence of the self-model's recursive grounding in Level 1 affect, rather than requiring it as an independent primitive.

Overall relationship: Complements (rather than "extends," the classification given to theories like COI, AST, and PToC that are less formally specified versions of RSP's insights). The distinction is that R2R provides something RSP genuinely lacks (perspectival entrapment as a candidate explanatory concept and an evolutionary story extending into pre-neural systems) rather than being a subset of RSP's framework. The relationship is symmetric: each provides what the other lacks. R2R and RSP represent convergence on the same core insight, consciousness as recursive self-modeling grounded in biological affect, from theoretical biology and hierarchical predictive coding respectively. The convergence strengthens both theories: if two methodologically distinct frameworks, starting from different premises and using different methods, arrive at structurally similar identity claims, the shared conclusion gains theoretical consilience beyond what either framework provides alone.

K. Synthesis

The following table summarizes how RSP relates to the major consciousness theories across five key dimensions:

Theory	Core Mechanism	Phenomenal Character	Developmental Account	Mathematical Formalism	RSP Relationship
GWT/GNW	Global broadcasting/ignition	Access only, not phenomenal	None	Partial (ignition dynamics)	Complements: RSP specifies broadcast content
SPA	Semantic pointers in neural populations	Via binding	Partial (learning)	Yes (NEF framework)	Complements: SPA as implementation framework
HOT	Higher-order meta-representation	Via meta-cognition	None	None	Extends: RSP adds recursion + affect grounding
FOR	First-order sensory representation	Asserted, not explained	None	None	Extends: RSP explains how FOR becomes conscious
COI	Cross-order integration	Via integration process	None	None	Extends: RSP formalizes and operationalizes the integration
AST	Attention schema via social cognition	Partial (attention only)	None	None	Extends: RSP incorporates AST + adds 4 schemas + recursion + formalism
SOMA	Learned metarepresentation via representational redescription	Via learning	None	None	Extends: RSP adds formalism + affect + dissolution
IIT	Integrated information (Φ)	Yes (qualia space)	None	Yes (Φ calculus)	Contradicts: architecture-specific vs. architecture-neutral
ART	Bottom-up/top-down resonance	Partial	Yes (learning)	Yes (resonance equations)	Complements: resonance as RSP implementation
CSS	3D consciousness space	Described, not explained	Partial (core $\rightarrow$ extended)	None	Extends: RSP provides causal

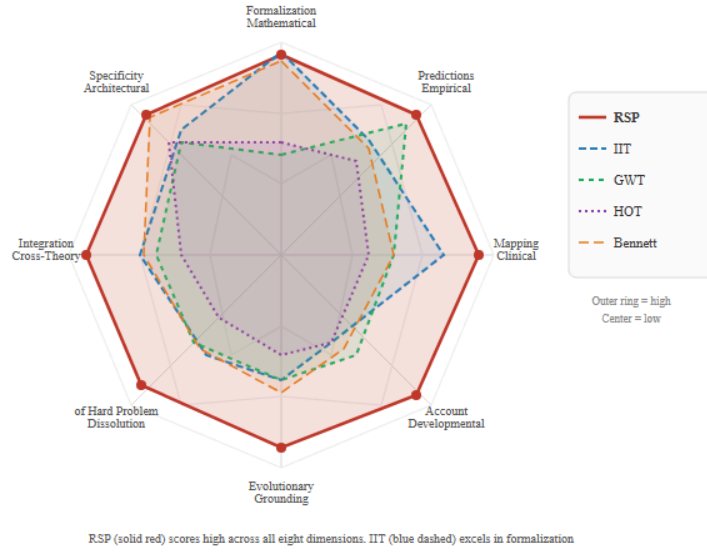
Theory	Core Mechanism	Phenomenal Character	Developmental Account	Mathematical Formalism	RSP Relationship
					mechanism for CSS's descriptive space
PToC	Tripartite hierarchy (Cons1-2-3)	Described	Yes (hierarchical)	None	Extends: RSP formalizes and neurobiologizes PToC's hierarchy
CP	Subcortical brainstem simulation	None	None	None	Complements: CP as RSP Level 1 substrate
RPT	Recurrent cortical processing	Via recurrence	None	None	Complements: RSP specifies recurrent content
Orch OR	Quantum collapse in microtubules	Via non-computation	None	Yes (quantum mechanics)	Contradicts: classical vs. quantum explanation
Basic Emotions	Innate discrete circuits (Ekman: expressions; Panksepp: subcortical systems)	Asserted	Innate (no development)	None	Partially contradicts: dimensional, not discrete (Panksepp's subcortical systems closer to RSP Level 1 than Ekman's expression model)
Constructed Emotion	Conceptual categorization of core affect	Constructed	Cultural learning	None	Extends: RSP adds schemas + recursive self-modeling to Barrett's core-affect-to-concept architecture
GT	Theme-field-margin phenomenology	Yes (descriptive)	None	Partial (dynamical systems)	Complements: RSP explains GT's structure

Theory	Core Mechanism	Phenomenal Character	Developmental Account	Mathematical Formalism	RSP Relationship
Stack Theory	W-maxing + tapestries of valence + orders of self	Yes (tapestries of valence)	Qualitative (1st→3rd order self)	Yes (set-theoretic proofs, experiments)	Complements: convergence on same architecture
R2R	Recursive depth + Value Saturation + perspectival entrapment	Yes (identity claim)	Phylogenetic (4 transitions)	None	Complements: convergence on identity claim

Key relationships:

- **RSP extends** COI, AST, CSS, and PToC by providing formal mechanisms for their core insights.
- **RSP complements** GWT/GNW, ART, CP, RPT, and GT by specifying content, mechanism, or substrate.
- **RSP extends** HOT, FOR, and SOMA by adding recursive self-modeling, affect grounding, and mathematical formalization.
- **RSP contradicts** IIT (on architecture-neutrality), Orch OR (on quantum necessity), and basic emotions theory (on discrete categories).
- **RSP complements** Bennett's Stack Theory: independent convergence on the same hierarchical, valence-grounded, recursively self-modeling architecture from computational learning theory rather than neuroscience.
- **RSP complements** Trukovich's R2R model: independent convergence on the identity claim (consciousness IS recursive self-modeling grounded in biological affect) from theoretical biology rather than predictive coding.

Figure 3. Theory Comparison Across Eight Dimensions



L. What RSP Uniquely Provides

Across the full range of consciousness theories surveyed in Sattin et al. (2021), RSP is the only framework that simultaneously offers:

1. **A specific mechanism for phenomenal character:** The recursive social prediction loop explains *why* consciousness feels like something, because the system predicts its own experience of predicting, creating a strange loop that is intrinsically first-personal.
2. **Multi-timescale architecture with explicit dynamics:** $\tau_1 \approx 100ms$, $\tau_2 \approx 1-10s$, $\tau_3 \approx 10s-minutes$, with formal timescale separation analysis (Appendix D).
3. **Four integrated mathematical formalizations unified by a fifth:** Hierarchical predictive coding (Appendix A), variational free energy (Appendix B), intrinsically motivated model-based RL (Appendix C), and cascade feedback control with contraction mapping convergence (Appendix D), unified via a master functional with proven cross-formalism equivalences (Appendix E). Appendix H extends the formalism to intelligence.
4. **A developmental progression grounded in mathematical necessity:** The Level $1 \rightarrow 2 \rightarrow 3$ construction order is not merely observed but required by the mathematics. Inner control loops must stabilize before outer loops can be commissioned (Appendix D), free energy at each level requires stability at the level below (Appendix B), and exploration policies require reward signals before they can improve (Appendix C).
5. **An interoceptive foundation:** Consciousness is always affectively toned because it is always grounded in core affect, the body's valence and arousal signals. No other theory in the survey gives

interoception this foundational role.

6. **Content specificity:** Unlike IIT (any integrated information) or GWT (any broadcast information), RSP specifies exactly *what* must be integrated: core affect + core schemas + recursive self-model. This specificity generates falsifiable predictions.
7. **A formal theory of intelligence grounded in the same mathematics as consciousness:** The master functional $\mathcal{J}$ simultaneously defines intelligence (Appendix H, Definition H.2) and constitutes phenomenal consciousness, establishing a constitutive identity between recursive social intelligence and consciousness at Level 3b. RSP provides a mathematically explicit account of intelligence derived from the same equations that formalize consciousness (Appendix H). The identity thesis, that recursive social intelligence and phenomenal consciousness are the same process, is not an informal claim but follows from the shared master functional $\mathcal{J}$ together with the architectural identity argument (Appendix H, §H.4). Bennett's (2025) Stack Theory convergently reaches the same conclusion from computational learning theory, and Trukovich's (2025a, 2025b) R2R model converges from theoretical biology. The three frameworks converge on the general thesis that consciousness is constituted by recursive self-modeling grounded in biological affect, though they differ on the specific mechanism (social prediction for RSP, w-maxing for Bennett, recursive depth for R2R). This consilience from three methodologically distinct starting points (hierarchical predictive coding, computational learning theory, and theoretical biology) strengthens the case for the shared core claim, even though the traditions are not fully independent.

No other theory in the scoping review combines all seven properties. Most offer one or two; several of the strongest theories (IIT, GNW, AST) offer three. RSP's integration of mechanism, mathematics, development, affect, content specificity, and intelligence theory in a single coherent framework represents its primary contribution to the science of consciousness.

XI. Conclusion: Beyond the Bat

What This Paper Has Accomplished

This paper has argued that the three-level RSP architecture is not merely one empirical description of how consciousness happens to work — it represents the conditions of possibility for the kind of consciousness we can coherently describe and investigate. The transcendental argument, grounded in convergent evidence from Kant, Jonas, Thompson, Merleau-Ponty, Husserl, Sartre, Strawson, and McDowell, shows that the ordering constraint (mattering before structure before recursion) is logically required given the framework's assumptions, while four independent mathematical formalisms all converge on the same three-level architecture. This is the paper's most distinctive philosophical contribution: not just a theory of consciousness, but an argument for why consciousness must be structured this way.

Three specific claims follow from this architecture:

First, the traditional assumption that phenomenal experience consists of "raw qualia"—meaning-independent, intrinsic phenomenal properties—rests on a conceptual confusion. Experience cannot be

separated from meaning any more than music can be separated from organized sound.

Second, conscious experiences are better understood as meaningful integrations: emergent products of the brain integrating biological core affect with culturally learned conceptual categories within predictive models. The three-level architecture (Level 1 core affect, Level 2 schemas, Level 3 recursive social prediction) is formalized across eight companion appendixes (A–H), with Level 3 further decomposed into conceptual categorization (3a) and the recursive self-modeling strange loop (3b) that constitutes phenomenal consciousness. The phenomenology doesn't exist at any single level but emerges from their synthesis.

Third, this reconceptualization dissolves rather than solves certain philosophical puzzles. Questions like "What is it like to be a bat?" or "Do we all see the same red?" assume something that doesn't exist—pure, meaning-independent phenomenology. These questions need reframing, not answering.

The New Research Program

Moving beyond the bat question opens productive research directions:

Neuroscience: Map how core affect systems interact with conceptual networks. Identify neural signatures of meaningful integration. Understand how predictions shape phenomenology at the cellular level.

Development: Track how emotional differentiation follows conceptual development. Study how social contexts enable or constrain emotional growth. Identify critical periods for emotional concept acquisition.

Cross-cultural psychology: Document which aspects of consciousness are universal (core affect) and which vary culturally (constructed emotions). Develop methods for describing culturally unique phenomenologies without assuming universal raw qualia.

Clinical psychology: Target meaningful integration mechanisms in therapy. Develop interventions that teach emotional concepts, modify predictions, or regulate core affect. Understand psychopathology as disrupted integration.

Comparative psychology: Investigate animal consciousness by studying their capacities for core affect, categorization, prediction, and self-modeling—not by trying to access their ineffable inner experience.

Disorders of consciousness: Map vegetative state, minimally conscious state, and locked-in syndrome onto RSP levels. Test whether TMS-EEG complexity (PCI) correlates with the number of active RSP levels. Compare propofol and ketamine effects on level-specific processing.

AI and consciousness: Design systems with appropriate functional architecture (valenced arousal, conceptual learning, predictive modeling) rather than trying to produce mysterious raw qualia.

Mathematical formalization: Eight companion appendixes (A–H) provide formal foundations for computational modeling and simulation of the RSP architecture. The predictive coding formulation (Appendix A) specifies the generative model and message-passing dynamics; the free energy formulation (Appendix B) yields gradient dynamics; the intrinsically motivated RL formulation (Appendix C) maps

onto existing computational frameworks; the control-theoretic formulation (Appendix D) enables stability analysis and prediction of pathological dynamics; the unified mathematical framework (Appendix E) proves that these four formalisms derive from a single master functional, identifies emergent properties (coordinate covariance, bounded error, precision as universal currency), and provides an algorithmic specification (§E.9) with phylogenetic parameterization showing how the architecture degrades gracefully from human to invertebrate consciousness. Appendix F consolidates the key equations; Appendix G maps clinical predictions onto the architecture; and Appendix H develops a formal theory of intelligence from the same master functional, establishing a constitutive identity between recursive social intelligence and consciousness at Level 3b. Future work should develop simulated RSP agents and test whether the predicted developmental sequence, psychopathological failure modes, and convergence properties emerge from the mathematics alone.

Philosophical Gains

Beyond empirical predictions, the meaningful integration framework clarifies philosophical issues:

The explanatory gap closes: Phenomenology is constitutive of recursive self-modeling, not a separate property that could be absent from it. The question "why does integration feel like something?" is dissolved as a category error, not merely narrowed to a more tractable form. Jackson's Mary gains new integration when she sees red for the first time; she does not access a previously hidden quale.

The problem of other minds dissolves: We don't need telepathic access to raw qualia but functional understanding of integration architecture.

The privacy of consciousness becomes intersubjectivity: Experience is neither purely private nor purely universal but intersubjectively structured through shared biology and culture.

Cultural variation becomes prediction: Different concepts create genuinely different phenomenologies—a central prediction, not puzzling exception.

Developmental emergence becomes intelligible: Consciousness builds gradually as concepts are acquired, not suddenly when raw qualia appear.

The Bat Reframed

Returning to Nagel's bat: it is now clear that "What is it like to be a bat?" misfires because it assumes bats have raw, meaning-independent experiences that we might access.

The better questions:

- What core affects do bats experience?
- What conceptual categories do bats possess?
- How do echolocation and flight shape bat predictions differently from human predictions?
- What aspects of bat consciousness are similar to ours (core affect, basic categorization) and what aspects differ (sensory modality, ecological niche)?

These questions are empirically tractable. We can study bat behavior, bat neuroscience, and bat ecology to understand bat meaningful integration. We don't need mystical access to ineffable bat-qualia.

The bat isn't a philosophical mystery demonstrating consciousness's irreducibility. It's an empirical challenge inviting comparative study of meaningful integration across species.

The Future

The RSP framework speaks directly to the question of artificial consciousness, perhaps the most urgent application of consciousness theory in the current moment. Under RSP, any candidate for consciousness must satisfy three architectural conditions: genuine interoceptive monitoring producing valence (Level 1), embodied schemas grounded in sensorimotor interaction (Level 2), and recursive social prediction converging to a stable self-model at depth

$$D_{max} \geq 3$$

(Level 3b). Current large language models satisfy none of these: they lack interoceptive monitoring, lack embodied schemas, and perform self-reference without recursive self-modeling. But RSP's conditions are functional, not biological. An artificial system that genuinely implemented the three-level architecture would, under the constitutive premise (§E.10), be a candidate for consciousness. This is a prediction of the theory, not a dismissal — and it is empirically addressable as architectures evolve.

Consciousness research stands at a crossroads. We can continue pursuing the raw qualia assumption, generating ever more elaborate attempts to explain the inexplicable. Or we can recognize that the assumption itself may be confused and pursue the more tractable project of understanding meaningful integration.

The meaningful integration framework dissolves the hard problem by showing that the question "why should integration feel like anything?" presupposes a separation between process and phenomenology that the RSP architecture rejects. It also dissolves numerous pseudoproblems, generates testable predictions, and connects consciousness research with broader cognitive science, developmental psychology, cultural anthropology, and affective neuroscience. The formal appendixes demonstrate that the RSP architecture is not merely a philosophical narrative but a mathematically specified dynamical system, one whose convergence can be proved, whose stability margins can be measured, whose failure modes map onto known psychopathologies, and whose developmental construction order follows necessarily from the cascade structure.

Most importantly, it suggests that consciousness is not a metaphysical anomaly but a natural phenomenon: complex, culturally embedded, and gradually evolved. Understanding consciousness requires understanding how biological systems create meaning through integration. That is a challenge, but it is the kind of challenge science can address.

The bat question should not haunt us as an unsolvable mystery. It should inspire us to better understand the functional architectures that enable meaningful integration across the remarkable diversity of conscious life.

Author's Note

This essay develops themes from an integrative framework on consciousness, emotion, and social development. The core insight—that feelings are meaningful integrations rather than raw qualia—emerged from attempts to reconcile seemingly contradictory findings across neuroscience, emotion theory, and developmental psychology. What began as an exercise in theoretical integration revealed a deeper philosophical issue: the assumption of raw qualia may be preventing progress in consciousness studies by generating pseudoproblems that obscure tractable empirical questions. If this paper succeeds in making that case, it will have served its purpose, not by solving the mystery of consciousness but by helping us ask better questions.

Methodological Note

This reference list follows APA 7th edition format with some adaptations for philosophical conventions. Where publication years are approximate or sources are unpublished manuscripts, this is indicated. For classical texts reprinted in modern editions, both original and reprint dates are provided (e.g., Locke 1689/1975).

References: Beyond the Bat - A Philosophical Essay on Consciousness

Complete Bibliography

A

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, 2(11), 1032-1037.

Alkire, M. T., Hudetz, A. G., & Tononi, G. (2008). Consciousness and anesthesia. *Science*, 322(5903), 876-880.

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology*, 5(4), 297-305.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MIT Press.

Behrmann, M., Avidan, G., Marotta, J. J., & Kimchi, R. (2005). Detailed exploration of face-related processing in congenital prosopagnosia: 1. Behavioral findings. *Journal of Cognitive Neuroscience*, 17(7), 1130-1149.

- Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1), 28-58.
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.
- Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41, 167-218.
- Barrett, L. F., Gross, J., Christensen, T. C., & Benvenuto, M. (2001). Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation. *Cognition and Emotion*, 15(6), 713-724.
-

B

- Bechara, A., Tranel, D., & Damasio, H. (2000). Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, 123(11), 2189-2202.
- Bennett, M. T. (2025). *How to build conscious machines*. Doctoral thesis, Australian National University. Preprint under review.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Broesch, T., Callaghan, T., Henrich, J., Murphy, C., & Rochat, P. (2011). Cultural variations in children's mirror self-recognition. *Journal of Cross-Cultural Psychology*, 42(6), 1018-1029.
- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. Guilford Press.
- Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, 24(10), 789-801.
- Blanke, O., Ortigue, S., Landis, T., & Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature*, 419(6904), 269-270.
- Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford University Press.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.
-

C

- Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, 12(5), 187-192.
- Chollet, F. (2019). On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.
- Connor, R. C., Smolker, R., & Bejder, L. (2006). Synchrony, social behaviour and alliance affiliation in Indian Ocean bottlenose dolphins, *Tursiops aduncus*. *Animal Behaviour*, 72(6), 1371-1378.

- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3), 201-215.
- Cacioppo, J. T., & Patrick, W. (2008). *Loneliness: Human nature and the need for social connection*. W. W. Norton & Company.
- Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press.
- Cambridge Declaration on Consciousness. (2012). Publicly proclaimed at the Francis Crick Memorial Conference on Consciousness in Human and non-Human Animals, Churchill College, University of Cambridge. <http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf>
- Camras, L. A. (1992). Expressive development and basic emotions. *Cognition and Emotion*, 6(3-4), 269-283.
- Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., ... & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200-219.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Chalmers, D. J. (2010). *The character of consciousness*. Oxford University Press.
- Chalmers, D. J. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, 25(9-10), 6-61.
- Chiao, J. Y., Harada, T., Komeda, H., Li, Z., Mano, Y., Saito, D., ... & Iidaka, T. (2008). Neural basis of individualistic and collectivistic views of self. *Human Brain Mapping*, 30(9), 2813-2820.
- Churchland, P. M. (1985). Reduction, qualia, and the direct introspection of brain states. *Journal of Philosophy*, 82(1), 8-28.
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.
- Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. *Behaviour Research and Therapy*, 58, 10-23.
- Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020). Learning to be conscious. *Trends in Cognitive Sciences*, 24(2), 112-123.
- Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern-day "wild child"*. Academic Press.

D

- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. Putnam.
- Descartes, R. (1641/1996). *Meditations on first philosophy* (J. Cottingham, Trans.). Cambridge University Press.
- Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Harcourt Brace.
- DeCasien, A. R., Williams, S. A., & Higham, J. P. (2017). Primate brain size is predicted by diet but not sociality. *Nature Ecology & Evolution*, 1(5), 0112.
- Della Sala, S., Marchetti, C., & Spinnler, H. (1991). Right-sided anarchic (alien) hand: A longitudinal study. *Neuropsychologia*, 29(11), 1113-1127.
- Decety, J., & Grèzes, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research*, 1079(1), 4-14.
- Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, 79(1-2), 1-37.
- Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200-227.
- Dozier, M., Stovall, K. C., Albus, K. E., & Bates, B. (2001). Attachment for infants in foster care: The role of caregiver state of mind. *Child Development*, 72(5), 1467-1477.
- Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, 44(4), 576-585.
- Dretske, F. (1995). *Naturalizing the mind*. MIT Press.
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178-190.
-

E

- Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.
- Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, 306(5703), 1903-1907.
- Ellis, H. D., & Young, A. W. (1990). Accounting for delusional misidentifications. *British Journal of Psychiatry*, 157(2), 239-248.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169-200.
-

F

- Finn, J. K., Tregenza, T., & Norman, M. D. (2009). Defensive tool use in a coconut-carrying octopus. *Current Biology*, 19(23), R1069-R1070.
- Fiorito, G., & Scotto, P. (1992). Observational learning in *Octopus vulgaris*. *Science*, 256(5056), 545-547.
- Fichte, J. G. (1794/1982). *Science of Knowledge* (P. Heath & J. Lachs, Trans.). Cambridge University Press.
- Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12), 11-39.
- Frith, U. (2003). *Autism: Explaining the enigma* (2nd ed.). Blackwell Publishing.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
- Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4), 531-534.
-

G

- Gallagher, S. (2005). *How the body shapes the mind*. Oxford University Press.
- Grossberg, S. (2021). *Conscious mind, resonant brain: How each brain makes a mind*. Oxford University Press.
- Gurwitsch, A. (1964). *The field of consciousness*. Duquesne University Press.
- Gallup, G. G., Jr. (1970). Chimpanzees: Self-recognition. *Science*, 167(3914), 86-87.
- Gennaro, R. J. (2012). *The consciousness paradox: Consciousness, concepts, and higher-order thoughts*. MIT Press.
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014a). Cultural relativity in perceiving emotion from vocalizations. *Psychological Science*, 25(4), 911-920.
- Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014b). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14(2), 251-262.
- Gerrans, P. (2000). Refining the explanation of Cotard's delusion. *Mind & Language*, 15(1), 111-122.
- Giacino, J. T., Fins, J. J., Laureys, S., & Schiff, N. D. (2014). Disorders of consciousness after acquired brain injury: The state of the science. *Nature Reviews Neurology*, 10(2), 99-114.
- Graziano, M. S. A. (2013). *Consciousness and the social brain*. Oxford University Press.
- Greenberg, L. S. (2002). *Emotion-focused therapy: Coaching clients to work through their feelings*. American Psychological Association.
-

H

- Hampton, R. R. (2001). Rhesus monkeys know when they remember. *Proceedings of the National Academy of Sciences*, 98(9), 5359-5362.
- Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, 59(4), 771-785.
- Hutter, M. (2005). *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer.
- Husserl, E. (1913/1983). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy, First Book* (F. Kersten, Trans.). Kluwer.
- Harlow, H. F. (1958). The nature of love. *American Psychologist*, 13(12), 673-685.
- Harlow, H. F., & Harlow, M. K. (1965). The affectional systems. In A. M. Schrier, H. F. Harlow, & F. Stollnitz (Eds.), *Behavior of nonhuman primates* (Vol. 2, pp. 287-334). Academic Press.
- Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour Research and Therapy*, 44(1), 1-25.
- Hirstein, W., & Ramachandran, V. S. (1997). Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proceedings of the Royal Society of London B*, 264(1380), 437-444.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Basic Books.
- Hofstadter, D. R. (2007). *I am a strange loop*. Basic Books.
- Hume, D. (1739/1978). *A treatise of human nature*. Oxford University Press.
-

I

- Itard, J. M. G. (1801/1962). *The wild boy of Aveyron* (G. Humphrey & M. Humphrey, Trans.). Appleton-Century-Crofts.
-

J

- Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. *Journal of Applied Psychology*, 95(1), 54-78.
- Jonas, H. (1966). *The phenomenon of life: Toward a philosophical biology*. Northwestern University Press.
- Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, 32(127), 127-136.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19), 7241-7244.

James, W. (1890). *The principles of psychology*. Henry Holt and Company.

K

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition*, 89(1), 25-41.

Kinderman, P., Dunbar, R. I. M., & Bentall, R. P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, 89(2), 191-204.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, 354(6308), 110-114.

Kant, I. (1781/1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press.

Kabat-Zinn, J. (1990). *Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness*. Delta.

Kappen, H. J. (2005). Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics*, P11011.

Kashdan, T. B., Barrett, L. F., & McKnight, P. E. (2015). Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. *Current Directions in Psychological Science*, 24(1), 10-16.

Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT Press.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.

Kohda, M., Hotta, T., Takeyama, T., Awata, S., Tanaka, H., Asai, J., & Jordan, A. L. (2019). If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals? *PLOS Biology*, 17(2), e3000021.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, 17(5), 307-321.

Koluchová, J. (1972). Severe deprivation in twins: A case study. *Journal of Child Psychology and Psychiatry*, 13(2), 107-114.

Koluchová, J. (1976). The further development of twins after severe and prolonged deprivation: A second report. *Journal of Child Psychology and Psychiatry*, 17(3), 181-188.

Kriegel, U. (2009). *Subjective consciousness: A self-representational theory*. Oxford University Press.

Kriegel, U. (Ed.). (2019). *The Oxford handbook of the philosophy of consciousness*. Oxford University Press.

L

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4), 391-444.

Light, G. A., Swerdlow, N. R., & Braff, D. L. (2007). Preattentive sensory processing as indexed by the MMN and P3a brain responses is associated with cognitive and psychosocial functioning in healthy adults. *Journal of Cognitive Neuroscience*, 19(10), 1624-1632.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Basic Books.

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

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8), 365-373.

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron*, 73(4), 653-676.

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

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. Oxford University Press.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., & Baron-Cohen, S. (2010). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage*, 56(3), 1832-1838.

Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv:1805.00909*.

Lindquist, K. A., & Gendron, M. (2013). What's in a word? Language constructs emotion perception. *Emotion Review*, 5(1), 66-71.

Locke, J. (1689/1975). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford University Press.

M

Mather, J. A. (2008). Cephalopod consciousness: Behavioural evidence. *Consciousness and Cognition*, 17(1), 37-48.

- McDowell, J. (1994). *Mind and world*. Harvard University Press.
- Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., ... & Koch, C. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *Nature*, 625, 373-381.
- Merleau-Ponty, M. (1945/2012). *Phenomenology of perception* (D. Landes, Trans.). Routledge.
- Della Sala, S. (1998). Disentangling the alien and anarchic hand. *Cognitive Neuropsychiatry*, 3(3), 191-207.
- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, 35(2), 63-78.
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(1), 63-81.
- Medford, N., Sierra, M., Baker, D., & David, A. S. (2005). Understanding and treating depersonalisation disorder. *Advances in Psychiatric Treatment*, 11(2), 92-100.
- Melzack, R. (1990). Phantom limbs and the concept of a neuromatrix. *Trends in Neurosciences*, 13(3), 88-92.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.
- Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of “pure” consciousness. *Philosophy and the Mind Sciences*, 1(1), 1-44.
-

N

- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nelson, C. A., Fox, N. A., & Zeanah, C. H. (2014). *Romania's abandoned children: Deprivation, brain development, and the struggle for recovery*. Harvard University Press.
-

O

- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology*, 10(5), e1003588.
- Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, 313(5792), 1402.
-

P

- Prigatano, G. P. (2010). *The study of anosognosia*. Oxford University Press.
- Prior, H., Schwarz, A., & Guentuerken, O. (2008). Mirror-induced behavior in the magpie (*Pica pica*): Evidence of self-recognition. *PLOS Biology*, 6(8), e202.
- Plotnik, J. M., de Waal, F. B. M., & Reiss, D. (2006). Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences*, 103(45), 17053-17057.
- Putnam, F. W. (1989). *Diagnosis and treatment of multiple personality disorder*. Guilford Press.
- Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. Oxford University Press.
- Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.
- Penrose, R., & Hameroff, S. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation*, 40(3-4), 453-480.
- Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford University Press.
-

R

- Reiss, D., & Marino, L. (2001). Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10), 5937-5942.
- Ramachandran, V. S. (1996). The evolutionary biology of self-deception, laughter, dreaming and depression: Some clues from anosognosia. *Medical Hypotheses*, 47(5), 347-362.
- Ramachandran, V. S., & Hirstein, W. (1998). The perception of phantom limbs: The D. O. Hebb lecture. *Brain*, 121(9), 1603-1630.
- Reinders, A. A. T. S., Nijenhuis, E. R. S., Paans, A. M. J., Korf, J., Willemsen, A. T. M., & den Boer, J. A. (2003). One brain, two selves. *NeuroImage*, 20(4), 2119-2125.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169-192.
- Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition*, 12(4), 717-731.
- Rochat, P. (2009). *Others in mind: Social origins of self-consciousness*. Cambridge University Press.
- Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford University Press.
- Rosenthal, D. M. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Guzeldere (Eds.), *The nature of consciousness* (pp. 729-753). MIT Press.
- Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin*, 110(3), 426-450.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.

S

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, 19(4), 1835-1842.

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, 29(1), 93-104.

Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for 'theory of mind.' *Brain*, 124(2), 279-286.

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., ... & Massimini, M. (2015). Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine. *Current Biology*, 25(23), 3099-3105.

Sartre, J.-P. (1943/1956). *Being and nothingness* (H. Barnes, Trans.). Philosophical Library.

Sierra, M., & Berrios, G. E. (1998). Depersonalization: Neurobiological perspectives. *Biological Psychiatry*, 44(9), 898-908.

Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565-573.

Singh, J. A. L., & Zingg, R. M. (1942). *Wolf-children and feral man*. Harper & Brothers.

Sattin, D., Bhatt, M., Gipponi, S., Leonardi, M., & Bhagwat, N. (2021). Theoretical models of consciousness: A scoping review. *Brain Sciences*, 11(5), 535.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.

Shoemaker, S. (2007). *Physical realization*. Oxford University Press.

Stern, D. N. (1985). *The interpersonal world of the infant: A view from psychoanalysis and developmental psychology*. Basic Books.

T

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

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

- Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.
- Tye, M. (2000). *Consciousness, color, and content*. MIT Press.
- Todorov, E. (2008). General duality between optimal control and estimation. *47th IEEE Conference on Decision and Control*, 4286-4292.
- Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, 215(3), 216-242.
- Trukovich, J. J. (2025a). From reactions to reflection: A recursive framework for the evolution of cognition and complexity. *BioSystems*, 250, 105402. <https://doi.org/10.1016/j.biosystems.2025.105402>
- Trukovich, J. J. (2025b). From reaction to reflection: A recursive framework for the evolution and structure of intelligence. *BioSystems*, 256, 105549. <https://doi.org/10.1016/j.biosystems.2025.105549>
- Tsai, J. L. (2007). Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science*, 2(3), 242-259.
-

U

- Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.
-

W

- Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2), 1-37.
- Wang, Q., & Fivush, R. (2005). Mother-child conversations of emotionally salient events: Exploring the functions of emotional reminiscing in European-American and Chinese families. *Social Development*, 14(3), 473-495.
- Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.
- Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, 72(3), 655-684.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004), 686-688.
- Widen, S. C., & Russell, J. A. (2008). Children acquire emotion categories gradually. *Cognitive Development*, 23(2), 291-312.
- Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.

Wittgenstein, L. (1953/2009). *Philosophical investigations* (G.E.M. Anscombe, P.M.S. Hacker, & J. Schulte, Trans., 4th ed.). Wiley-Blackwell.

Y

Young, A. W., & Leafhead, K. M. (1996). Betwixt life and death: Case studies of the Cotard delusion. In P. W. Halligan & J. C. Marshall (Eds.), *Method in madness: Case studies in cognitive neuropsychiatry* (pp. 147-171). Psychology Press.

Yagmurlu, B., Berument, S. K., & Celimli, S. (2005). The role of institution and home contexts in theory of mind development. *Journal of Applied Developmental Psychology*, 26(5), 521-537.

Z

Yoshimi, J., & Vinson, D. W. (2015). Extending Gurwitsch's field theory of consciousness. *Consciousness and Cognition*, 34, 104-123.

Zahavi, D. (2014). *Self and other: Exploring subjectivity, empathy, and shame*. Oxford University Press.

S (continued)

Shanon, B. (2008). A psychological theory of consciousness. *Journal of Consciousness Studies*, 15(5), 5-47.

Appendix A: The RSP as a Hierarchical Predictive Coding System

Overview

This appendix formalizes the Recursive Social Prediction (RSP) model as a hierarchical predictive coding system. We specify the generative model at each level, the message-passing dynamics between levels, and the precision-weighting mechanisms that govern attention and learning. The formulation builds on Rao & Ballard (1999), Friston (2005, 2010), and Clark (2013), extending hierarchical predictive coding to encompass affect, schemas, and recursive self-modeling.

A.1 Notation and Definitions

A.1.1 Core Variables

Symbol	Domain	Definition
t	$\mathbb{N}$	Discrete time index
$\mathbf{x}(t)$	$\mathbb{R}^{n_x}$	Exteroceptive sensory input vector (visual, auditory, tactile, etc.)
$\mathbf{i}(t)$	$\mathbb{R}^{n_i}$	Interoceptive input vector (visceral, proprioceptive, nociceptive signals)
$v(t)$	$[-1, 1]$	Valence dimension of core affect (unpleasant $\rightarrow$ pleasant)
$\alpha(t)$	$[0, 1]$	Arousal dimension of core affect (calm $\rightarrow$ activated)
$\mathbf{a}(t)$	$[-1, 1] \times [0, 1]$	Core affect vector $\mathbf{a}(t) = (v(t), \alpha(t))$
$\mathcal{S}$	—	Set of core schemas $\{s_{\text{body}}, s_{\text{spatial}}, s_{\text{aff-hom}}, s_{\text{attn}}, s_{\text{self}}\}$
$\mathbf{s}_k(t)$	$\mathbb{R}^{n_k}$	State vector of schema $k \in \mathcal{S}$
$\mathbf{z}(t)$	$\mathbb{R}^{n_z}$	Integrated schema representation (structured affect)
$\mathcal{C}$	—	Set of available conceptual categories (culturally learned)
$c^*(t)$	$\mathcal{C}$	Selected concept maximizing posterior probability
$\mathbf{e}(t)$	$\mathbb{R}^{n_e}$	Constituted experience (meaningful integration of $\mathbf{z}(t)$ with $c^*(t)$)
$\mathcal{M}(t)$	$\mathbb{R}^{n_m}$	Self-model state vector at time t
$\mathcal{O}$	—	Set of other-agent models $\{o_1, o_2, \dots, o_K\}$
$o_j(t)$	$\mathbb{R}^{n_o}$	Model of other agent j 's mental state
$\hat{\mathcal{M}}_j(t)$	$\mathbb{R}^{n_m}$	Predicted self-as-seen-by agent j
$\hat{\mathbf{e}}(t+1)$	$\mathbb{R}^{n_e}$	Predicted future experience
$\boldsymbol{\varepsilon}_l(t)$	$\mathbb{R}^{n_l}$	Prediction error at level l
$\boldsymbol{\Pi}_l(t)$	$\mathbb{R}^{n_l \times n_l}$	Precision matrix at level l (inverse covariance of prediction errors)
$\boldsymbol{\Sigma}_l$	$\mathbb{R}^{n_l \times n_l}$	Temporal transition noise covariance at level l
$\boldsymbol{\eta}_l(t)$	$\mathbb{R}^{n_l}$	Temporal prediction error: $\boldsymbol{\mu}_l(t) - f_l(\boldsymbol{\mu}_l(t-1), \boldsymbol{\mu}_{l+1}(t-1))$
$\boldsymbol{\xi}_l(t)$	$\mathbb{R}^{n_l}$	Precision-weighted prediction error: $\boldsymbol{\Pi}_l \cdot \boldsymbol{\varepsilon}_l$
$\Phi(t)$	—	Phenomenal state: the integrated tuple $\langle \mathbf{e}(t), \mathcal{M}(t), \hat{\mathbf{e}}(t+1) \rangle$

A.1.2 Level Indices

Level	Label	Generative Model	Primary Substrate
$l = 1$	Core Affect	$g_1(\boldsymbol{\mu}_2)$: predicts interoceptive signals from schema states	Brainstem, hypothalamus, amygdala, anterior insula
$l = 2$	Core Schemas	$g_2(\boldsymbol{\mu}_3)$: predicts schema configurations from conceptual/self-model states	Somatosensory, hippocampal, frontoparietal, TPJ circuits
$l = 3$	Recursive Social Prediction	$g_3(\boldsymbol{\mu}_3^{\text{rec}})$: predicts experience from recursive self-model	mPFC, TPJ/STS, dlPFC, default mode network

A.1.3 Operators

Operator	Definition
$g_l(\cdot)$	Generative function at level l : maps higher-level representations to predictions about the level below
$f_l(\cdot)$	State transition function at level l : governs temporal dynamics of level- l representations
$\boldsymbol{\mu}_l(t)$	Sufficient statistics (expectations) of the approximate posterior at level l
$\dot{\boldsymbol{\mu}}_l(t)$	Rate of change of expectations at level l (gradient descent on free energy)
D	Temporal derivative operator: $D\boldsymbol{\mu} = \dot{\boldsymbol{\mu}}$

A.2 The Generative Model

A.2.1 Hierarchical Structure

The RSP generative model is a three-level hierarchy. Each level l generates predictions about the level below via a nonlinear mapping g_l , with additive Gaussian noise scaled by precision $\boldsymbol{\Pi}_l$:

$$\text{Level 1: } \mathbf{i}(t) = g_1(\boldsymbol{\mu}_2(t)) + \boldsymbol{\Pi}_1^{-1/2} \boldsymbol{\omega}_1$$

$$\text{Level 2: } \boldsymbol{\mu}_1(t) = g_2(\boldsymbol{\mu}_3(t)) + \boldsymbol{\Pi}_2^{-1/2} \boldsymbol{\omega}_2$$

$$\text{Level 3: } \boldsymbol{\mu}_2(t) = g_3(\boldsymbol{\mu}_3^{\text{rec}}(t)) + \boldsymbol{\Pi}_3^{-1/2} \boldsymbol{\omega}_3$$

where $\boldsymbol{\omega}_l \sim \mathcal{N}(0, \mathbf{I})$ and $\boldsymbol{\mu}_3^{\text{rec}}$ denotes the recursively updated self-model (see §A.4).

A.2.2 Prediction Errors

At each level, prediction errors encode the discrepancy between actual input and top-down prediction:

$$\begin{aligned} \boldsymbol{\varepsilon}_1(t) &= \mathbf{i}(t) - g_1(\boldsymbol{\mu}_2(t)) \\ \boldsymbol{\varepsilon}_2(t) &= \boldsymbol{\mu}_1(t) - g_2(\boldsymbol{\mu}_3(t)) \end{aligned}$$

$$\varepsilon_3(t) = \boldsymbol{\mu}_2(t) - g_3(\boldsymbol{\mu}_3^{\text{rec}}(t))$$

These errors are precision-weighted before propagating upward:

$$\boldsymbol{\xi}_l(t) = \boldsymbol{\Pi}_l(t) \cdot \varepsilon_l(t)$$

where $\boldsymbol{\xi}_l$ is the precision-weighted prediction error at level l .

A.2.3 Temporal Dynamics

Each level also generates predictions about its own future state via a state transition model:

$$\boldsymbol{\mu}_l(t+1) = f_l(\boldsymbol{\mu}_l(t), \boldsymbol{\mu}_{l+1}(t)) + \boldsymbol{\Sigma}_l^{1/2} \boldsymbol{\nu}_l$$

where $\boldsymbol{\Sigma}_l$ is the temporal transition noise covariance and $\boldsymbol{\nu}_l \sim \mathcal{N}(0, \mathbf{I})$.

A.3 Level-Specific Generative Functions

A.3.1 Level 1: Core Affect Generation

The Level 1 generative function maps schema-level expectations to predicted interoceptive states:

$$g_1(\boldsymbol{\mu}_2) = \mathbf{W}_1^{\text{aff}} \sigma(\mathbf{z}(t)) + \mathbf{b}_1$$

where:

- $\mathbf{W}_1^{\text{aff}} \in \mathbb{R}^{n_i \times n_z}$ is the learned mapping from integrated schemas to interoceptive predictions
- $\sigma(\cdot)$ is a nonlinear activation function (e.g., sigmoid or tanh)
- $\mathbf{b}_1$ is a bias term encoding homeostatic set-points

The core affect vector is derived from the precision-weighted prediction error:

$$\mathbf{a}(t) = \phi_{\text{aff}}(\boldsymbol{\xi}_1(t))$$

where

$$\phi_{\text{aff}} : \mathbb{R}^{n_i} \rightarrow [-1, 1] \times [0, 1]$$

is a learned projection onto the valence-arousal plane. Interoceptive prediction errors that signal homeostatic deviation produce negative valence and high arousal; errors consistent with allostasis produce positive valence.

A.3.2 Level 2: Schema Integration

Each schema $k \in \mathcal{S}$ maintains its own state vector $\mathbf{s}_k(t)$ updated by both bottom-up sensory input and top-down predictions from Level 3:

$$\mathbf{s}_k(t) = f_k^{\text{schema}}(\mathbf{s}_k(t-1), \mathbf{x}_k(t), \mathbf{a}(t), g_{2,k}(\boldsymbol{\mu}_3(t)))$$

where:

- $\mathbf{x}_k(t)$ is the sensory input relevant to schema k (e.g., proprioceptive signals for body schema, spatial cues for spatial schema)
- $g_{2,k}$ is the Level 3 $\rightarrow$ schema- k prediction function
- f_k^{schema} is the schema-specific update function (learned)

The five schemas and their primary inputs:

Schema k	Primary Input $\mathbf{X}_k$	State $\mathbf{S}_k$ Encodes
s_{body}	Proprioceptive, tactile, vestibular	Body configuration, boundaries, ownership
s_{spatial}	Visual-spatial, hippocampal place/grid cells	Allocentric and egocentric spatial maps
$s_{\text{aff-hom}}$	Interoceptive, hypothalamic	Homeostatic state, energy balance, drive states
s_{attn}	Frontoparietal priority maps	Attentional focus, salience, awareness attribution
s_{self}	Mirror system, TPJ, precuneus	Self-other distinction, agency, ownership

Schema integration combines individual schema states into a unified structured representation:

$$\mathbf{z}(t) = \text{Integrate}(\mathbf{a}(t), \{\mathbf{s}_k(t)\}_{k \in \mathcal{S}}) = \sigma(\mathbf{W}_{\text{int}} \cdot \text{concat}(\mathbf{a}(t), \mathbf{s}_1(t), \dots, \mathbf{s}_5(t)) + \mathbf{b}_{\text{int}})$$

where

$$\mathbf{W}_{\text{int}}$$

is a learned integration matrix implementing cross-schema binding.

A.3.3 Level 3: Recursive Social Prediction

Level 3 operates in two stages:

Stage 3a — Conceptual categorization:

$$c^*(t) = \arg \max_{c \in \mathcal{C}} P(c \mid \mathbf{z}(t), \text{context}(t))$$

The constituted experience is the integration of structured affect with the selected concept:

$$\mathbf{e}(t) = \phi_{\text{exp}}(\mathbf{z}(t), \mathbf{h}_{c^*})$$

where $\mathbf{h}_{c^*} \in \mathbb{R}^{n_c}$ is the learned embedding of concept c^* .

Stage 3b — Recursive self-modeling (see §A.4 for full specification).

A.4 The Recursive Self-Model

A.4.1 Self-Model Update

The self-model

$$\mathcal{M}(t)$$

is updated by integrating the current constituted experience:

$$\mathcal{M}(t) = f_{\mathcal{M}}(\mathcal{M}(t-1), \mathbf{e}(t)) = \mathcal{M}(t-1) + \lambda_{\mathcal{M}} \cdot \tanh(\mathbf{W}_{\mathcal{M}} \cdot [\mathcal{M}(t-1); \mathbf{e}(t)] + \mathbf{b}_{\mathcal{M}})$$

where $\lambda_{\mathcal{M}} \in (0, 1]$ is the self-model learning rate and $[\cdot; \cdot]$ denotes concatenation.

A.4.2 Other-Model Maintenance

For each agent j in the social context $\mathcal{O}$:

$$o_j(t) = f_o(o_j(t-1), \mathbf{x}_j^{\text{social}}(t))$$

where $\mathbf{x}_j^{\text{social}}(t)$ encodes observed behavior, facial expressions, vocal prosody, and inferred intentions of agent j .

A.4.3 The Strange Loop: Recursive Depth

For recursion depth

$$d = 1, \dots, D_{\max}$$

:

$$\hat{\mathcal{M}}_j^{(d)}(t) = \text{PredictSelfAsSeenBy}(\mathcal{M}^{(d-1)}(t), o_j(t))$$

$$\hat{\mathcal{M}}^{(d)}(t) = \sum_{j=1}^{|\mathcal{O}|} w_j \cdot \hat{\mathcal{M}}_j^{(d)}(t)$$

$$\mathcal{M}^{(d)}(t) = (1 - \beta_d) \cdot \mathcal{M}^{(d-1)}(t) + \beta_d \cdot \hat{\mathcal{M}}^{(d)}(t)$$

where:

- $\mathcal{M}^{(0)}(t) = \mathcal{M}(t)$ after the Stage 3a update
- $w_j \in [0, 1]$ are salience weights for each other-agent ($\sum_j w_j = 1$)
- $\beta_d \in (0, 1)$ is a depth-dependent mixing coefficient, typically $\beta_d = \beta_0 \cdot \gamma^d$ with decay $\gamma \in (0, 1)$, ensuring deeper recursion has diminishing influence

The final self-model after

$$D_{\max}$$

recursion steps is:

$$\boldsymbol{\mu}_3^{\text{rec}}(t) = \mathcal{M}^{(D_{\max})}(t)$$

A.4.4 Recursion Depth and Convergence

The recursion terminates when either

$$d = D_{\max}$$

or the update magnitude falls below threshold:

$$\|\mathcal{M}^{(d)}(t) - \mathcal{M}^{(d-1)}(t)\| < \delta_{\text{conv}}$$

Typical biological values: $D_{\max} \approx 3\text{--}5$ (corresponding to the depth of recursive mentalizing observed empirically: "I think that you think that I think..."), $\delta_{\text{conv}} \approx 0.01$.

A.5 Message Passing Dynamics

A.5.1 Bottom-Up: Precision-Weighted Prediction Errors

Prediction errors propagate upward, weighted by precision:

$$\xi_l(t) = \Pi_l(t) \cdot \varepsilon_l(t) \quad (\text{upward message from level } l)$$

A.5.2 Top-Down: Predictions

Predictions propagate downward via the generative functions:

$$\hat{\mu}_{l-1}(t) = g_l(\mu_l(t)) \quad (\text{downward message from level } l)$$

A.5.3 Expectation Update (Gradient Descent on Free Energy)

At each level, expectations are updated by gradient descent on variational free energy (see Appendix B):

$$\dot{\mu}_l = D\mu_l - \kappa_l \cdot \frac{\partial F}{\partial \mu_l}$$

Expanding the free energy gradient (omitting temporal terms for clarity; see Appendix B, §B.4.1 for the full expansion):

$$\frac{\partial F}{\partial \mu_l} = -\frac{\partial g_{l-1}^T}{\partial \mu_l} \xi_{l-1} + \xi_{l+1}$$

where $g_{l-1}(\mu_l)$ is the generative function that takes μ_l as input to predict the level below:

- The first term drives expectations toward explaining bottom-up prediction errors (via the Jacobian of the generative function that maps from level l)
- The second term ensures consistency with top-down predictions from the level above (μ_l enters linearly in ε_{l+1} , so no Jacobian is needed)
- $\kappa_l > 0$ is the learning rate at level l

A.5.4 Precision Update

Precisions are themselves updated to minimize free energy:

$$\dot{\mathbf{\Pi}}_l = -\eta_l \cdot \frac{\partial F}{\partial \mathbf{\Pi}_l} = \eta_l \left(\frac{1}{2} \mathbf{\Pi}_l^{-1} - \frac{1}{2} \boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l^T \right)$$

This implements **attention**: high precision at a level means prediction errors there are weighted more heavily, driving the system to explain those errors. The attention schema (s_{attn}) modulates $\mathbf{\Pi}_l$ across levels, implementing the RSP model’s claim that attention determines which computations reach consciousness.

A.6 The Phenomenal State

With the dynamics of message passing, expectation update, and precision modulation in hand, we can now characterize what the system’s state looks like when these dynamics reach (approximate) convergence — the phenomenal state.

The phenomenal state

$$\Phi(t)$$

is not a separate output but the integrated state of the hierarchy at convergence:

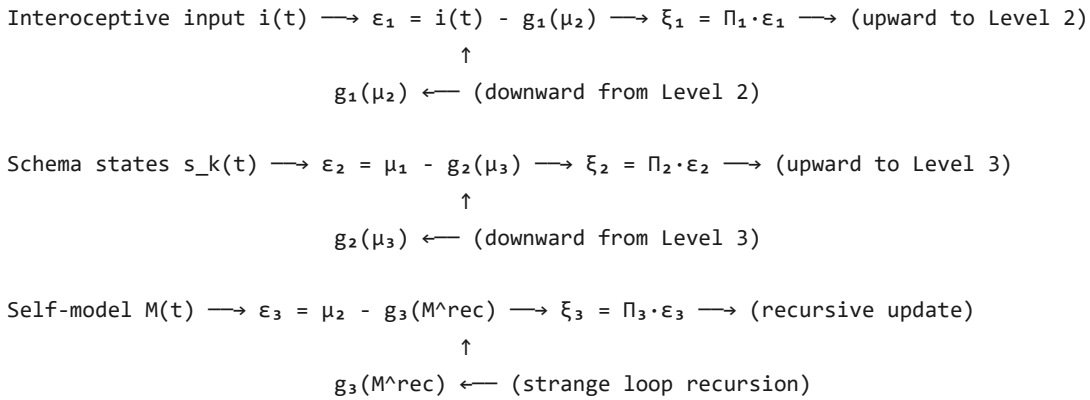
$$\Phi(t) = \langle \mathbf{e}(t), \mathcal{M}^{(D_{\max})}(t), \hat{\mathbf{e}}(t+1) \rangle$$

where:

- $\mathbf{e}(t)$ is the constituted experience (what it feels like)
- $\mathcal{M}^{(D_{\max})}(t)$ is the recursively refined self-model (who is feeling it)
- $\hat{\mathbf{e}}(t+1)$ is the predicted next experience (the forward-looking character of consciousness)

Consciousness, on this account, is the process of hierarchical inference itself — not a separate property generated by the inference, but the inference as experienced from within the system performing it.

A.7 Summary of Information Flow



Each level simultaneously:

1. Receives precision-weighted prediction errors from below (surprise signal)
2. Generates predictions downward (expectation signal)
3. Updates its own expectations to minimize free energy
4. Updates its precision estimates (attention allocation)

This bidirectional flow, operating continuously, constitutes the stream of consciousness.

References for Appendix A

- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181-204.
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815-836.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
- Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1), 79-87.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565-573.

Appendix B: Free Energy Formulation of RSP

Overview

This appendix derives the variational free energy functional for the Recursive Social Prediction (RSP) model. We show how each level of the hierarchy contributes to a single scalar free energy bound, how minimizing this bound implements perception, learning, and attention, and how the recursive self-model enters as a uniquely self-referential term. The formulation builds on Friston (2005, 2010), Buckley et al. (2017), and Bogacz (2017), extending variational inference to encompass affect, schemas, and recursive social prediction.

Notation. All symbols follow Appendix A. In particular: μ_l denotes expectations at level l ; Π_l denotes precision matrices; $g_l(\cdot)$ denotes generative functions; $\varepsilon_l = \text{input}_l - g_l(\mu_{l+1})$ denotes prediction errors.

B.1 The Variational Free Energy Principle

B.1.1 From Bayes to Free Energy

An organism maintains a generative model $p(\mathbf{y}, \boldsymbol{\theta})$ of sensory observations $\mathbf{y}$ and hidden causes $\boldsymbol{\theta}$. Exact Bayesian inference — computing the posterior $p(\boldsymbol{\theta} \mid \mathbf{y})$ — is intractable for hierarchical nonlinear models. Instead, the system maintains an approximate posterior $q(\boldsymbol{\theta})$ and minimizes the divergence between this approximation and the true posterior.

The **variational free energy** is:

$$F = -\ln p(\mathbf{y}) + D_{\text{KL}}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \mathbf{y})]$$

Since

$$D_{\text{KL}} \geq 0$$

, free energy is an upper bound on surprisal:

$$F \geq -\ln p(\mathbf{y}) = \text{surprisal}$$

Minimizing F with respect to q tightens this bound, making the approximate posterior approach the true posterior. Minimizing F with respect to action changes sensory input to reduce surprise.

B.1.2 The Laplace Approximation

Under the Laplace (or "maximum a posteriori") approximation, the approximate posterior is a Gaussian centered on its mode:

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ are the sufficient statistics (expectations) and $\boldsymbol{\Sigma}$ is the posterior covariance. Under this approximation, the free energy simplifies to an expression involving only prediction errors and precisions (see §B.2).

B.1.3 Definitions

Symbol	Domain	Definition
F	$\mathbb{R}$	Total variational free energy (scalar)
F_l	$\mathbb{R}$	Free energy contribution from level l
$\mathbf{y}$	$\mathbb{R}^{n_y}$	Complete sensory observation vector $\mathbf{y} = [\mathbf{x}(t); \mathbf{i}(t)]$
$\boldsymbol{\theta}$	—	All hidden causes: $\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\mu}_3^{\text{rec}}\}$
$q(\boldsymbol{\theta})$	—	Approximate posterior (recognition density)
$p(\mathbf{y}, \boldsymbol{\theta})$	—	Generative model (joint density over observations and causes)
$D_{\text{KL}}[\cdot \cdot]$	$\mathbb{R}_{\geq 0}$	Kullback-Leibler divergence
$\mathcal{L}$	$\mathbb{R}$	Negative free energy (evidence lower bound, ELBO): $\mathcal{L} = -F$

B.2 Free Energy for the RSP Hierarchy

B.2.1 Total Free Energy Decomposition

Under the Laplace approximation with Gaussian noise at each level (as specified in Appendix A, §A.2.1), the total free energy decomposes into a sum of level-specific contributions:

$$F = F_1 + F_2 + F_3 + F_{\text{temporal}}$$

Each level contributes a term proportional to the precision-weighted squared prediction error plus a complexity penalty:

$$F_l = \underbrace{\frac{1}{2} \boldsymbol{\varepsilon}_l^T \boldsymbol{\Pi}_l \boldsymbol{\varepsilon}_l}_{\text{accuracy}} - \underbrace{\frac{1}{2} \ln |\boldsymbol{\Pi}_l|}_{\text{complexity}} + \underbrace{\frac{n_l}{2} \ln(2\pi)}_{\text{constant}}$$

where:

- $\boldsymbol{\varepsilon}_l(t) = \text{input}_l(t) - g_l(\boldsymbol{\mu}_{l+1}(t))$ is the prediction error at level l
- $\boldsymbol{\Pi}_l(t)$ is the precision (inverse covariance) of prediction errors
- $n_l = \dim(\boldsymbol{\varepsilon}_l)$ is the dimensionality of level l
- The **accuracy** term penalizes prediction errors proportional to confidence (precision)
- The **complexity** term penalizes excessive confidence, preventing overfitting

B.2.2 Temporal Free Energy

The temporal component penalizes deviations from predicted state transitions:

$$F_{\text{temporal}} = \sum_{l=1}^3 \frac{1}{2} \boldsymbol{\eta}_l^T \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\eta}_l - \frac{1}{2} \ln |\boldsymbol{\Sigma}_l^{-1}| + \frac{n_l}{2} \ln(2\pi)$$

where

$$\boldsymbol{\eta}_l(t) = \boldsymbol{\mu}_l(t) - f_l(\boldsymbol{\mu}_l(t-1), \boldsymbol{\mu}_{l+1}(t-1))$$

is the temporal prediction error — the discrepancy between the current state and what the transition model predicted.

B.3 Level-Specific Free Energy

B.3.1 Level 1: Core Affect Free Energy

$$F_1 = \frac{1}{2} [\mathbf{i}(t) - g_1(\boldsymbol{\mu}_2)]^T \boldsymbol{\Pi}_1 [\mathbf{i}(t) - g_1(\boldsymbol{\mu}_2)] - \frac{1}{2} \ln |\boldsymbol{\Pi}_1|$$

Interpretation. Level 1 free energy is high when interoceptive signals deviate from what the body schema predicts. A racing heart in a safe context produces large ϵ_1 , driving up F_1 , which the system resolves by either updating schema expectations (perception: "I must be anxious") or changing the body state (action: deep breathing). High $\boldsymbol{\Pi}_1$ means the system trusts interoceptive signals — interoceptive sensitivity that Seth (2013) links to emotional awareness.

Affective valence as free energy gradient. The core affect vector $\mathbf{a}(t) = \phi_{\text{aff}}(\boldsymbol{\xi}_1(t))$ can be rewritten in terms of the free energy gradient:

$$v(t) \approx -\text{sign} \left(\frac{\partial F_1}{\partial \boldsymbol{\mu}_2} \right) \cdot \left\| \frac{\partial F_1}{\partial \boldsymbol{\mu}_2} \right\|$$

Negative valence tracks the direction and magnitude of free energy increase — felt unpleasantness is the phenomenology of rising interoceptive surprise.

B.3.2 Level 2: Schema Free Energy

$$F_2 = \frac{1}{2} [\boldsymbol{\mu}_1 - g_2(\boldsymbol{\mu}_3)]^T \boldsymbol{\Pi}_2 [\boldsymbol{\mu}_1 - g_2(\boldsymbol{\mu}_3)] - \frac{1}{2} \ln |\boldsymbol{\Pi}_2|$$

The schema-level free energy has a richer structure because $\boldsymbol{\mu}_1$ is itself a vector of affect-enriched states and g_2 must predict five schema configurations simultaneously. Expanding:

$$F_2 = \sum_{k \in \mathcal{S}} F_{2,k} + F_{2,\text{cross}}$$

where:

$$F_{2,k} = \frac{1}{2} [\mathbf{s}_k(t) - g_{2,k}(\boldsymbol{\mu}_3)]^T \boldsymbol{\Pi}_{2,k} [\mathbf{s}_k(t) - g_{2,k}(\boldsymbol{\mu}_3)]$$

is the free energy for schema k , and $F_{2,\text{cross}}$ captures cross-schema coupling (off-diagonal blocks of $\boldsymbol{\Pi}_2$). The five schema-specific terms:

Schema	$F_{2,k}$ Measures Surprise About...
s_{body}	Body configuration deviating from predicted posture/ownership
s_{spatial}	Spatial location deviating from predicted navigation state
$s_{\text{aff-hom}}$	Homeostatic balance deviating from predicted metabolic state
s_{attn}	Attentional focus deviating from predicted salience map
s_{self}	Self-other boundary deviating from predicted agency/ownership

Attention as precision optimization. The attention schema s_{attn} has a special role: it modulates $\boldsymbol{\Pi}_{2,k}$ for all other schemas. This means the attention schema’s state determines how much weight each schema’s prediction errors carry in the total free energy. Formally:

$$\boldsymbol{\Pi}_{2,k}(t) = \boldsymbol{\Pi}_{2,k}^{\text{base}} \cdot \sigma(\mathbf{W}_{\text{attn},k} \cdot \mathbf{s}_{\text{attn}}(t))$$

where

$$\sigma$$

is a sigmoid ensuring positive definite precision. This implements the RSP model’s claim that consciousness is determined by which computations receive attentional precision-weighting.

B.3.3 Level 3: Recursive Self-Model Free Energy

$$F_3 = \frac{1}{2} [\boldsymbol{\mu}_2 - g_3(\boldsymbol{\mu}_3^{\text{rec}})]^T \boldsymbol{\Pi}_3 [\boldsymbol{\mu}_2 - g_3(\boldsymbol{\mu}_3^{\text{rec}})] - \frac{1}{2} \ln |\boldsymbol{\Pi}_3|$$

The self-referential term. Unlike Levels 1 and 2, the hidden cause at Level 3 is the recursively updated self-model $\boldsymbol{\mu}_3^{\text{rec}} = \mathcal{M}^{(D_{\text{max}})}(t)$. This means the system minimizes surprise about its own schema states *as predicted by a model of itself as seen through others’ eyes*. The free energy is not merely about matching observations to predictions — it is about the self-model’s coherence with its social reflections.

Expanding the recursive structure (from Appendix A, §A.4.3):

$$F_3^{\text{rec}} = \sum_{d=1}^{D_{\text{max}}} \gamma^d \cdot \frac{1}{2} \|\mathcal{M}^{(d)}(t) - \mathcal{M}^{(d-1)}(t)\|_{\boldsymbol{\Pi}_3^{(d)}}^2$$

where:

- $\gamma \in (0, 1)$ is the recursion discount factor (deeper recursions contribute less)
- $\boldsymbol{\Pi}_3^{(d)}$ is the precision at recursion depth d

- $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^T \mathbf{A} \mathbf{x}$ is the Mahalanobis norm

This sum over recursion depths formalizes the "strange loop": each depth adds a term penalizing the discrepancy between how I see myself and how I predict others see me seeing myself... to depth

$$D_{\max}$$

.

B.4 Gradient Dynamics: Minimizing Free Energy

B.4.1 Expectation Updates

The expectation dynamics at each level follow gradient descent on the total free energy (as stated in Appendix A, §A.5.3):

$$\dot{\boldsymbol{\mu}}_l = D\boldsymbol{\mu}_l - \kappa_l \cdot \frac{\partial F}{\partial \boldsymbol{\mu}_l}$$

We now derive the update direction

$$-\frac{\partial F}{\partial \boldsymbol{\mu}_l}$$

explicitly for each level (the negative gradient, which is the direction expectations move under gradient descent).

Level 1:

$$-\frac{\partial F}{\partial \boldsymbol{\mu}_1} = -\boldsymbol{\Pi}_2 \cdot \boldsymbol{\varepsilon}_2 - \boldsymbol{\Sigma}_1^{-1} \cdot \boldsymbol{\eta}_1$$

The first term drives Level 1 expectations toward the Level 2 prediction (reducing prediction error

$$\boldsymbol{\varepsilon}_2 = \boldsymbol{\mu}_1 - g_2(\boldsymbol{\mu}_3)$$

). The second term enforces temporal consistency (how much Level 1 deviates from its own predicted dynamics).

Level 2:

$$-\frac{\partial F}{\partial \boldsymbol{\mu}_2} = \underbrace{\frac{\partial g_1^T}{\partial \boldsymbol{\mu}_2} \boldsymbol{\Pi}_1 \boldsymbol{\varepsilon}_1}_{\text{explain interoceptive errors}} - \underbrace{\boldsymbol{\Pi}_3 \boldsymbol{\varepsilon}_3}_{\text{reduce self-model error}} - \underbrace{\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\eta}_2}_{\text{temporal consistency}}$$

Level 2 expectations are simultaneously pulled by three forces:

1. **Downward:** adjust schemas to better predict interoceptive signals (via g_1)

2. **Upward**: reduce prediction error at Level 3 (since $\varepsilon_3 = \mu_2 - g_3(\mu_3^{\text{rec}})$, the system adjusts μ_2 toward the self-model's prediction)
3. **Temporal**: remain consistent with the transition model's predictions

This multi-directional gradient is the formal expression of the "meaningful integration" that the RSP model claims constitutes experience.

Level 3:

$$-\frac{\partial F}{\partial \mu_3} = \frac{\partial g_2^T}{\partial \mu_3} \Pi_2 \varepsilon_2 - \frac{\partial F_3^{\text{rec}}}{\partial \mu_3} - \Sigma_3^{-1} \eta_3$$

The recursive term

$$\frac{\partial F_3^{\text{rec}}}{\partial \mu_3}$$

is what makes Level 3 uniquely self-referential: the self-model's update depends on the self-model itself, creating a fixed-point problem that the system solves iteratively (the strange loop).

B.4.2 Precision Updates

Precisions are updated by gradient descent on F with respect to Π_l (Appendix A, §A.5.4):

$$\dot{\Pi}_l = -\eta_l \cdot \frac{\partial F}{\partial \Pi_l} = \eta_l \left(\frac{1}{2} \Pi_l^{-1} - \frac{1}{2} \varepsilon_l \varepsilon_l^T \right)$$

At equilibrium (

$$\dot{\Pi}_l = 0$$

):

$$\Pi_l^* = \langle \varepsilon_l \varepsilon_l^T \rangle_t^{-1}$$

where $\langle \cdot \rangle_t$ denotes time-averaging (the instantaneous outer product $\varepsilon_l \varepsilon_l^T$ is rank-1; the time-averaged empirical covariance is full-rank and invertible). This means the optimal precision equals the inverse of the empirical error covariance — the system learns to trust levels whose prediction errors are small and stable. This is a self-tuning mechanism: levels that predict well gain more influence over the hierarchy.

B.4.3 Action as Free Energy Minimization

The organism can also minimize free energy through action — changing sensory input rather than updating internal states:

$$\dot{\mathbf{u}} = -\kappa_u \cdot \frac{\partial F}{\partial \mathbf{u}} = -\kappa_u \cdot \frac{\partial \mathbf{y}^T}{\partial \mathbf{u}} \Pi_1 \varepsilon_1$$

where $\mathbf{u}(t)$ is the motor command vector and $\frac{\partial \mathbf{y}}{\partial \mathbf{u}}$ is the (learned) mapping from actions to sensory consequences.

Active inference in the RSP context. The organism doesn't just passively update beliefs — it acts to make its predictions come true. Emotional expression (facial, vocal, postural) is active inference on the interoceptive model: the system predicts a calm body state and acts to bring the body into alignment. Social behavior is active inference on the self-model: the system predicts how others see it and acts to bring social reality into alignment with its self-concept.

B.5 Consciousness as Free Energy Minimization

B.5.1 The Phenomenal State Revisited

The phenomenal state $\Phi(t) = \langle \mathbf{e}(t), \mathcal{M}^{(D_{\max})}(t), \hat{\mathbf{e}}(t+1) \rangle$ is the state of the hierarchy at the point where free energy is (locally) minimized. Consciousness, on this account, is not a reward for minimizing free energy — it is the process of minimizing free energy across all three levels simultaneously.

B.5.2 Why Three Levels Are Necessary

The total free energy

$$F = F_1 + F_2 + F_3 + F_{\text{temporal}}$$

requires all three levels for full minimization:

Level Absent	Consequence	Phenomenological Analog
No Level 1	F_1 unconstrained: no interoceptive grounding	Depersonalization — experience without felt body
No Level 2	F_2 unconstrained: no schema integration	Dreamless sleep — raw affect without structure
No Level 3	F_3 unconstrained: no self-model	Minimal consciousness — structured experience without self-awareness
All three	Full minimization: coherent self-aware experience	Normal waking consciousness

B.5.3 The Hard Problem as a Free Energy Problem

The "explanatory gap" between physical processes and subjective experience corresponds, in this framework, to the gap between:

1. **The objective description:** gradient descent on a variational bound, $\dot{\boldsymbol{\mu}}_l = D\boldsymbol{\mu}_l - \kappa_l \frac{\partial F}{\partial \boldsymbol{\mu}_l}$
2. **The subjective description:** "what it is like" to be the system performing this inference

The RSP model does not claim to close this gap by reduction. Instead, it dissolves the gap by showing that the objective description already has the structure we attribute to subjective experience:

- **Valence** = direction of the free energy gradient at Level 1
- **Intentionality** = the generative model's predictions (content is what is predicted)
- **Unity** = the single scalar F integrating all levels
- **Selfhood** = the recursive fixed point μ_3^{rec}
- **Temporal flow** = the temporal prediction errors η_t and predicted future $\hat{e}(t + 1)$

These are not merely analogies. If one accepts the functionalist premise that consciousness is constituted by its relational and structural properties — a substantive philosophical position defended in the paper body (§VIII) — then these correspondences are structural identities: the mathematical properties of free energy minimization in a three-level recursive hierarchy *are* the properties we require of consciousness.

B.6 Learning: Long-Term Free Energy Minimization

The preceding analysis characterizes consciousness at a single moment — the phenomenal state at convergence of the free energy dynamics. We now turn to how the free energy landscape itself changes over time through learning, asking how the generative model's parameters and structure are shaped by experience.

B.6.1 Parameter Learning

The generative model parameters

$$\theta = \{\mathbf{W}_l, \mathbf{b}_l, \mathbf{W}_{\text{int}}, \mathbf{W}_{\mathcal{M}}\}$$

are updated on a slower timescale by gradient descent on expected free energy:

$$\dot{\theta} = -\kappa_{\theta} \cdot \left\langle \frac{\partial F}{\partial \theta} \right\rangle_{\tau}$$

where $\langle \cdot \rangle_{\tau}$ denotes averaging over a time window τ (minutes to hours). This corresponds to synaptic plasticity — the slow adjustment of connection weights that shapes the generative model.

B.6.2 Model Selection and Cultural Learning

Over developmental timescales (months to years), the system selects among competing generative model structures by minimizing the expected free energy of the model itself:

$$m^* = \arg \min_m \mathbb{E}[F_m]$$

This implements the RSP model's claim that conceptual categories

$\mathcal{C}$

are culturally learned: different cultures provide different generative models for categorizing affect, and the child adopts the model that best predicts their social environment’s responses. The free energy framework explains why emotional concepts differ across cultures — each culture’s model is locally optimal for its particular pattern of social contingencies.

B.6.3 Hierarchical Learning Rates

A key feature of the RSP free energy landscape is that learning rates differ by level:

Level	Timescale	Learning Rate	What Changes
Level 1	Milliseconds–seconds	$\kappa_1 \gg \kappa_2 \gg \kappa_3$	Interoceptive expectations (fast)
Level 2	Seconds–minutes	κ_2	Schema configurations (moderate)
Level 3	Minutes–hours	κ_3	Self-model, emotional concepts (slow)
Parameters	Hours–years	κ_θ	Synaptic weights (very slow)
Model structure	Years–lifetime	—	Conceptual repertoire (developmental)

This hierarchy of timescales ensures stability: fast lower-level updates don’t destabilize slow upper-level representations, while upper-level changes propagate gradually downward through the generative model. Disruptions to this timescale separation (e.g., trauma rapidly rewriting the self-model) produce pathological states.

B.7 Expected Free Energy and Planning

B.7.1 Active Inference with Temporal Depth

The system doesn’t just minimize current free energy — it minimizes *expected* free energy over a planning horizon T :

$$G(\pi) = \sum_{\tau=t}^{t+T} \mathbb{E}_{q(\mathbf{y}_\tau, \boldsymbol{\theta}|\pi)} [F(\mathbf{y}_\tau, \boldsymbol{\theta})]$$

where

$$\pi$$

is a policy (sequence of planned actions) and the expectation is over predicted future observations and states under that policy. This decomposes into:

$$G(\pi) = \underbrace{- \sum_{\tau} \mathbb{E}_q [\ln p(\mathbf{y}_\tau)]}_{\text{pragmatic value (reward)}} - \underbrace{\sum_{\tau} \mathbb{E}_q [D_{\text{KL}}[q(\boldsymbol{\theta}|\mathbf{y}_\tau) \| q(\boldsymbol{\theta})]]}_{\text{epistemic value (information gain)}}$$

The first term drives the system toward predicted observations that are consistent with its preferences (homeostasis, positive valence). The second term drives epistemic foraging — seeking observations that reduce uncertainty about hidden states.

B.7.2 Social Prediction as Expected Free Energy

At Level 3, the planning horizon includes predicted social outcomes. The expected free energy of a social action

$$\pi_{\text{social}}$$

is:

$$G_3(\pi_{\text{social}}) = \mathbb{E} \left[\sum_{d=1}^{D_{\max}} \gamma^d \left\| \mathcal{M}^{(d)}(t + \tau) - \mathcal{M}^{(d-1)}(t + \tau) \right\|_{\Pi_3^{(d)}}^2 \middle| \pi_{\text{social}} \right]$$

The system chooses social actions that minimize predicted discrepancy between self-model and socially-reflected self-model. This is the free energy formulation of impression management, self-presentation, and social conformity — all driven by the same principle that drives perception and action: minimizing surprise.

B.8 Summary: The Free Energy Landscape of Consciousness

The total free energy functional for the RSP model:

$$F = \underbrace{\frac{1}{2} \boldsymbol{\varepsilon}_1^T \boldsymbol{\Pi}_1 \boldsymbol{\varepsilon}_1}_{F_1: \text{interoceptive}} + \underbrace{\frac{1}{2} \boldsymbol{\varepsilon}_2^T \boldsymbol{\Pi}_2 \boldsymbol{\varepsilon}_2}_{F_2: \text{schema}} + \underbrace{\frac{1}{2} \boldsymbol{\varepsilon}_3^T \boldsymbol{\Pi}_3 \boldsymbol{\varepsilon}_3 + F_3^{\text{rec}}}_{F_3: \text{recursive self}} + \underbrace{\sum_l \frac{1}{2} \boldsymbol{\eta}_l^T \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\eta}_l}_{F_{\text{temp}}: \text{temporal}} - \underbrace{\frac{1}{2} \sum_l \ln |\boldsymbol{\Pi}_l|}_{C: \text{complexity}}$$

The system minimizes

$$F$$

simultaneously through:

1. **Perception:** updating $\boldsymbol{\mu}_l$ (changing beliefs to match observations)
2. **Action:** updating $\mathbf{u}$ (changing observations to match beliefs)
3. **Attention:** updating $\boldsymbol{\Pi}_l$ (changing confidence to match reliability)
4. **Learning:** updating $\boldsymbol{\theta}$ (changing the generative model)
5. **Planning:** selecting π (choosing actions to minimize expected future F)

Consciousness is the state of the hierarchy when all five optimization processes operate together — the felt sense of being an organism that knows itself, predicts its social world, and acts to maintain coherence between its self-model and its place in that world.

References for Appendix B

- Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, 76, 198-211.
- Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, 81, 55-79.
- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B*, 360(1456), 815-836.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, 29(1), 1-49.
- Parr, T., & Friston, K. J. (2019). Generalised free energy and active inference. *Biological Cybernetics*, 113(5-6), 495-513.
- Seth, A. K. (2013). Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11), 565-573.

Appendix C: Intrinsically Motivated Reinforcement Learning in RSP

Overview

This appendix recasts the Recursive Social Prediction (RSP) model in the language of reinforcement learning (RL). We show that valence functions as a reward signal, schemas as learned state representations, and Level 3 recursive prediction as model-based planning. The formulation bridges affective neuroscience and computational RL, revealing that the RSP hierarchy implements a form of **unsupervised model-based reinforcement learning** where the agent learns both the reward function and the state representation without external supervision.

Notation. All symbols from Appendix A are retained. New RL-specific notation is introduced in §C.1.

C.1 RL Notation and Mapping

C.1.1 New Symbols

Symbol	Domain	Definition
$r(t)$	$\mathbb{R}$	Scalar reward signal at time t
$\mathbf{s}^{\text{RL}}(t)$	$\mathbb{R}^{n_s}$	RL state vector (agent’s representation of its situation)
$\pi(\mathbf{u} \mid \mathbf{s}^{\text{RL}})$	$[0, 1]$	Policy: probability of action $\mathbf{u}$ given state $\mathbf{s}^{\text{RL}}$
$V(\mathbf{s}^{\text{RL}})$	$\mathbb{R}$	Value function: expected cumulative future reward from state $\mathbf{s}^{\text{RL}}$
$Q(\mathbf{s}^{\text{RL}}, \mathbf{u})$	$\mathbb{R}$	Action-value function: expected return from taking action $\mathbf{u}$ in state $\mathbf{s}^{\text{RL}}$
$\delta(t)$	$\mathbb{R}$	Temporal difference (TD) error
γ_{RL}	$(0, 1)$	Discount factor for future rewards
$\mathcal{T}(\mathbf{s}' \mid \mathbf{s}^{\text{RL}}, \mathbf{u})$	$[0, 1]$	Transition model: predicted next state given current state and action
$\mathcal{R}(\mathbf{s}^{\text{RL}})$	$\mathbb{R}$	Reward model: predicted reward from state $\mathbf{s}^{\text{RL}}$

C.1.2 RSP-to-RL Mapping

RL Concept	RSP Equivalent	Explanation
Reward $r(t)$	Valence $v(t)$	Interoceptive prediction errors, signed by homeostatic direction
State $\mathbf{s}^{\text{RL}}(t)$	Schema integration $\mathbf{z}(t)$	Five-schema structured representation of the agent's situation
Action $\mathbf{u}(t)$	Motor command + social behavior	Both body-directed and socially-directed actions
Transition model $\mathcal{T}$	Level 3 generative model g_3	Predicts how schemas change given self-model and actions
Reward model $\mathcal{R}$	Level 1 generative model g_1	Predicts interoceptive (valence) consequences of schema states
Model-based planning	Strange loop recursion	Simulating social consequences before acting
Exploration	Epistemic foraging	Seeking states that reduce uncertainty about self and others
Value function V	Expected free energy $-G$	Negated expected free energy (states with low G are high-value)

C.2 Valence as Reward

C.2.1 The Reward Signal

In standard RL, reward is supplied externally. In the RSP model, reward is generated internally through interoceptive prediction:

$$r(t) = v(t) = \phi_v(\xi_1(t))$$

where

$$\phi_v : \mathbb{R}^{n_i} \rightarrow [-1, 1]$$

projects the precision-weighted interoceptive prediction error onto the valence dimension. This projection is learned, not innate — but its inputs (interoceptive signals) are phylogenetically constrained.

Why this is unsupervised. No external agent specifies what counts as rewarding. The reward function emerges from the structure of the generative model:

- Predictions consistent with homeostasis $\rightarrow$ small $\varepsilon_1 \rightarrow$ low $F_1 \rightarrow$ positive valence
- Predictions violated by homeostatic disruption $\rightarrow$ large $\varepsilon_1 \rightarrow$ high $F_1 \rightarrow$ negative valence

The organism learns to associate certain schema configurations with positive or negative valence, creating a self-supervised reward landscape.

C.2.2 Reward Decomposition

The scalar reward decomposes into contributions from each schema:

$$r(t) = \sum_{k \in \mathcal{S}} w_k^r \cdot r_k(t) + r_{\text{cross}}(t)$$

where:

$$r_k(t) = -\frac{1}{2} \boldsymbol{\varepsilon}_{1,k}^T \boldsymbol{\Pi}_{1,k} \boldsymbol{\varepsilon}_{1,k}$$

is the schema- k contribution to reward (negative because large prediction errors reduce reward), w_k^r are learned reward weights, and r_{cross} captures cross-schema interactions.

Schema	Reward Contribution	Example
s_{body}	Body-state comfort, pain absence	Warmth, satiation, physical ease
s_{spatial}	Environmental safety, navigational coherence	Familiar surroundings, orientation
$s_{\text{aff-hom}}$	Homeostatic balance	Blood sugar stability, temperature regulation
s_{attn}	Cognitive engagement, flow	Focused attention, low distraction
s_{self}	Agency, social belonging	Competence, autonomy, relatedness

C.2.3 Dopaminergic Implementation

The temporal difference error in the RSP framework is:

$$\delta(t) = r(t+1) + \gamma_{\text{RL}} \cdot V(\mathbf{z}(t+1)) - V(\mathbf{z}(t))$$

This maps onto phasic dopamine firing in the ventral tegmental area (VTA) and substantia nigra, which signals the difference between received and expected reward. The RSP model predicts that dopaminergic TD errors should correlate with *interoceptive* prediction errors specifically — not just with external rewards — consistent with evidence that visceral and interoceptive signals modulate dopaminergic processing (Critchley & Harrison, 2013).

C.3 Schemas as Learned State Representations

C.3.1 The Representation Learning Problem

In RL, the choice of state representation

$$\mathbf{s}^{\text{RL}}$$

determines what the agent can learn and plan about. A poor representation makes learning impossible; a good one makes it efficient. The RSP model’s five schemas are the agent’s learned state representation:

$$\mathbf{s}^{\text{RL}}(t) = \mathbf{z}(t) = \text{Integrate}(\mathbf{a}(t), \{\mathbf{s}_k(t)\}_{k \in \mathcal{S}})$$

This representation is learned through the same free energy minimization that drives perception and action (Appendix B). The schemas develop ontogenetically: body schema first (infancy), spatial schema next, affective homeostasis in parallel, attention as executive functions mature, and self-other distinction last (emerging with social cognition in toddlerhood).

C.3.2 Auxiliary Tasks and Schema Learning

Each schema can be viewed as solving an **auxiliary task** (Jaderberg et al., 2017) that provides useful gradients for the main RL objective:

Schema	Auxiliary Task	Learning Signal
s_{body}	Predict proprioceptive next-state	Body movement prediction error
s_{spatial}	Predict visual-spatial next-state	Navigation prediction error
$s_{\text{aff-hom}}$	Predict homeostatic next-state	Interoceptive prediction error
s_{attn}	Predict which inputs matter	Salience prediction error
s_{self}	Predict social consequences of own actions	Agency prediction error

Each auxiliary task provides its own prediction error signal, enabling the system to learn rich representations even when the primary reward signal

$$r(t)$$

is sparse. This is structurally identical to the use of auxiliary tasks in deep RL to accelerate representation learning.

C.3.3 State Abstraction

The integration step $\text{Integrate}(\cdot)$ performs state abstraction — compressing the high-dimensional sensory input into a compact representation sufficient for prediction and action selection. The integrated schema vector $\mathbf{z}(t)$ satisfies two properties:

1. **Sufficiency:** $\mathbf{z}(t)$ retains enough information to predict future rewards: $r(t + \tau) \perp \mathbf{y}(t) \mid \mathbf{z}(t)$ for all $\tau > 0$
2. **Minimality:** $\mathbf{z}(t)$ discards information irrelevant to reward prediction, implementing the information bottleneck principle (Tishby et al., 2000)

The attention schema

$$s_{\text{attn}}$$

determines what information passes through the bottleneck — implementing the RSP model’s claim that attention determines the contents of consciousness.

C.4 Level 3 as Model-Based Planning

C.4.1 Model-Based vs. Model-Free

Standard RL distinguishes:

- **Model-free:** learn value/policy directly from experience (habitual behavior)
- **Model-based:** learn a world model and plan by simulating future trajectories (deliberate behavior)

The RSP hierarchy implements both:

RL Type	RSP Level	Mechanism
Model-free	Level 1 → Level 2	Cached affect-to-action mappings (habits, conditioned responses)
Model-based	Level 3	Recursive simulation of social consequences
Hybrid	Cross-level	Model-based predictions guide model-free learning

C.4.2 Planning via Recursive Simulation

Level 3 planning proceeds by iteratively simulating:

Step 1. Predict the consequences of a candidate action $\mathbf{u}_{\text{cand}}$ on the schema state:

$$\hat{\mathbf{z}}(t+1 \mid \mathbf{u}_{\text{cand}}) = g_3^{\text{dyn}}(\mathbf{z}(t), \mathbf{u}_{\text{cand}}, \boldsymbol{\mu}_3^{\text{rec}}(t))$$

Step 2. Predict the reward (valence) of the resulting state:

$$\hat{r}(t+1 \mid \mathbf{u}_{\text{cand}}) = \mathcal{R}(\hat{\mathbf{z}}(t+1)) = -\frac{1}{2} \hat{\mathbf{e}}_1^T \boldsymbol{\Pi}_1 \hat{\mathbf{e}}_1$$

Step 3. Predict others' responses to the action (the social model):

$$\hat{o}_j(t+1 \mid \mathbf{u}_{\text{cand}}) = f_o^{\text{pred}}(o_j(t), \mathbf{u}_{\text{cand}})$$

Step 4. Predict the recursive self-model update given others' predicted responses:

$$\hat{\mathcal{M}}^{(d)}(t+1 \mid \mathbf{u}_{\text{cand}}) = \text{RecurseUpdate}(\hat{\mathcal{M}}^{(d-1)}, \{\hat{o}_j(t+1)\})$$

Step 5. Evaluate expected free energy of the candidate action:

$$G(\mathbf{u}_{\text{cand}}) = \hat{F}_1 + \hat{F}_2 + \hat{F}_3 + \hat{F}_3^{\text{rec}}$$

Step 6. Select the action minimizing expected free energy:

$$\mathbf{u}^*(t) = \arg \min_{\mathbf{u}_{\text{cand}}} G(\mathbf{u}_{\text{cand}})$$

This planning process is the computational implementation of "thinking before acting" — and specifically, "thinking about what others will think of me if I act this way."

C.4.3 The Social Reward Shaping

The recursive self-model introduces a social component to the reward function that goes beyond individual homeostasis:

$$r_{\text{social}}(t) = - \sum_{d=1}^{D_{\max}} \gamma^d \cdot \frac{1}{2} \left\| \mathcal{M}^{(d)}(t) - \mathcal{M}^{(d-1)}(t) \right\|_{\Pi_3^{(d)}}^2$$

This is a form of **reward shaping** (Ng et al., 1999): the base reward $r(t) = v(t)$ is augmented by a social consistency term that penalizes discrepancies between self-model and socially-reflected self-model. The shaped reward:

$$r_{\text{total}}(t) = r(t) + \lambda_{\text{social}} \cdot r_{\text{social}}(t)$$

ensures that the agent values not only homeostatic comfort but also social coherence — explaining why humans will endure physical discomfort (fasting, cold exposure, painful rituals) to maintain social standing.

C.5 The Intrinsically Motivated Learning Pipeline

The model-based planning of §C.4 depends on having a learned world model and a structured state representation. But how are these acquired? The answer is an unsupervised learning pipeline in which each level’s outputs bootstrap the next level’s learning — a process that must occur in a specific developmental order.

C.5.1 Three Phases of Learning

The RSP model’s learning can be understood as three nested RL problems, each operating on a different timescale:

Phase 1: Reward Function Learning (Level 1)

$$\hat{r}(t) = \phi_v(\Pi_1(t) \cdot [\mathbf{i}(t) - g_1(\boldsymbol{\mu}_2(t))])$$

The system learns what counts as rewarding by learning to predict its own interoceptive states. This is analogous to **intrinsic motivation** in RL (Schmidhuber, 1991; Oudeyer & Kaplan, 2007): the reward signal is generated internally through prediction error, not externally specified.

Timescale: Fast (milliseconds). Updated with each new interoceptive observation.

Phase 2: State Representation Learning (Level 2)

$$\mathbf{z}(t) = \text{Integrate}(\mathbf{a}(t), \{f_k^{\text{schema}}(\cdot)\}_{k \in \mathcal{S}})$$

The schemas learn to extract the information from sensory input that is relevant for predicting future reward. This is unsupervised representation learning — the schemas are shaped by the statistics of the sensory input and the structure of the reward landscape.

Timescale: Moderate (seconds to minutes for updates; months to years for structural development).

Phase 3: World Model and Policy Learning (Level 3)

$$\pi^*(\mathbf{u} \mid \mathbf{z}, \boldsymbol{\mu}_3^{\text{rec}}) = \arg \min_{\pi} G(\pi)$$

The system learns a transition model, a social model, and a policy — all by minimizing expected free energy. The world model includes not only physical dynamics but social dynamics (how others respond to one's actions), making this a multi-agent RL problem solved from the perspective of a single agent with theory of mind.

Timescale: Slow (minutes to hours for tactical planning; years for strategic self-model revision).

C.5.2 Why "Unsupervised"

The term "unsupervised" is justified on three grounds:

1. **No external reward specification.** The reward function $r(t) = v(t)$ is derived from the agent's own interoceptive prediction errors, not from an external signal.
2. **No external state labels.** The schemas $\{s_k\}$ are learned representations, not pre-specified features. The agent discovers which features of its sensory input are relevant for prediction and control.
3. **No external curriculum.** The developmental sequence (body schema $\rightarrow$ spatial $\rightarrow$ homeostatic $\rightarrow$ attentional $\rightarrow$ self-other) emerges from the structure of the learning problem itself — simpler predictions must be stable before more complex predictions can be learned atop them.

C.5.3 Comparison with Deep RL Architectures

Feature	Standard Deep RL	RSP RL
Reward	External, fixed	Internal, learned (interoceptive surprise)
State	Raw pixels or hand-crafted	Learned schemas (5 auxiliary tasks)
Model	Optional (model-based variants)	Required (3-level generative hierarchy)
Multi-agent	Separate module	Integrated via recursive self-model
Exploration	ϵ -greedy, curiosity bonus	Epistemic foraging via expected free energy
Value function	Scalar expected return	Expected free energy (multi-objective)
Hierarchy	Optional (options, feudal)	Required (3 levels, each with distinct timescale)

C.6 Temporal Difference Learning Across Levels

C.6.1 Level-Specific TD Errors

Each level of the hierarchy generates its own temporal difference error:

Level 1 TD (Affective TD):

$$\delta_1(t) = r(t+1) + \gamma_{\text{RL}} \cdot V_1(\mathbf{a}(t+1)) - V_1(\mathbf{a}(t))$$

Tracks whether affect is better or worse than expected. Positive δ_1 = pleasant surprise. Negative δ_1 = disappointment. Neural correlate: phasic dopamine in VTA.

Level 2 TD (Schema TD):

$$\delta_2(t) = r(t+1) + \gamma_{\text{RL}} \cdot V_2(\mathbf{z}(t+1)) - V_2(\mathbf{z}(t))$$

Tracks whether the structured situation is better or worse than predicted by the schema integration. Positive δ_2 = "things are going well." Negative δ_2 = "this situation is deteriorating." Neural correlate: dopaminergic projections to dorsal striatum.

Level 3 TD (Social TD):

$$\delta_3(t) = r_{\text{total}}(t+1) + \gamma_{\text{RL}} \cdot V_3(\boldsymbol{\mu}_3^{\text{rec}}(t+1)) - V_3(\boldsymbol{\mu}_3^{\text{rec}}(t))$$

Tracks whether the self-model's social prediction was confirmed or violated. Positive δ_3 = "others see me as I see myself" (social validation). Negative δ_3 = "I misjudged how others see me" (social surprise, embarrassment). Neural correlate: dopaminergic projections to mPFC, social reward processing.

C.6.2 Cross-Level TD Propagation

TD errors propagate across levels, creating a cascade:

$$\delta_1 \xrightarrow{\text{drives}} \text{schema update} \xrightarrow{\text{generates}} \delta_2 \xrightarrow{\text{drives}} \text{self-model update} \xrightarrow{\text{generates}} \delta_3$$

A sudden loud noise:

1. Level 1: $\delta_1 < 0$ (interoceptive surprise, negative valence)
2. Level 2: $\delta_2 < 0$ (spatial schema violated — something unexpected in environment)
3. Level 3: $\delta_3 \approx 0$ (self-model unaffected, unless in social context where startle response is embarrassing, then $\delta_3 < 0$)

This cascade formalizes how a simple sensory event can escalate into a social-emotional experience through the hierarchy.

C.7 Exploration, Curiosity, and the Epistemic Drive**C.7.1 Intrinsic Motivation**

The expected free energy (Appendix B, §B.7.1) naturally decomposes into pragmatic value (reward-seeking) and epistemic value (uncertainty-reducing):

$$G(\pi) = \underbrace{G_{\text{pragmatic}}(\pi)}_{\text{seek reward}} + \underbrace{G_{\text{epistemic}}(\pi)}_{\text{reduce uncertainty}}$$

The epistemic term provides an intrinsic motivation signal — curiosity — that drives the organism to explore states that resolve uncertainty about its generative model. At Level 3, this becomes specifically *social* curiosity:

$$G_{\text{epistemic}}^{\text{social}}(\pi) = \mathbb{E} \left[\sum_{j=1}^{|\mathcal{O}|} D_{\text{KL}} [q(o_j(t + \tau) \mid \mathbf{y}_{t+\tau}) \parallel q(o_j(t))] \mid \pi \right]$$

This is the expected information gain about other agents' mental states — the computational drive to understand others that manifests as social curiosity, gossip, and perspective-taking.

C.7.2 Exploration-Exploitation in Social Contexts

The exploration-exploitation trade-off takes a distinctive form at Level 3:

- **Exploitation:** Act in ways that maintain the current self-model and social relationships (conformity, consistency)
- **Exploration:** Act in novel ways that could update the self-model and reveal new social information (risk-taking, self-disclosure, challenging social norms)

The balance is governed by the precision of the self-model Π_3 : high self-model precision favors exploitation (rigid self-concept), low precision favors exploration (flexible self-concept). This connects to personality psychology: trait openness corresponds to low Π_3 , while trait conscientiousness corresponds to high Π_3 .

C.8 Summary: The RL Architecture of Consciousness

The RSP model implements a complete reinforcement learning system:

Level	Role	Key Equations	TD Error
LEVEL 3: Model-Based Planner	Plans by simulating social world and self	World model: $g_3(\text{social dynamics, self-model})$	δ_3 (social validation / surprise)
		Policy: $\pi^*(u \mid z, \mathcal{M}^{\text{rec}}) = \arg \min G(u)$	
		Social reward: $r_{\text{social}} = -\frac{1}{2} \ \mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\ _{\Pi_3^{(d)}}^2$	
LEVEL 2: State Representation Learner	Learns structured state representations	Schemas: $\{s_{\text{body}}, s_{\text{spatial}}, s_{\text{aff-hom}}, s_{\text{attn}}, s_{\text{self}}\}$	δ_2 (situation improving / deteriorating)
		Integration: $z(t) = \text{Integrate}(a(t), \{s_k(t)\})$	
		Auxiliary tasks: 5 prediction errors $\rightarrow$ gradients	
LEVEL 1: Reward Function Generator	Generates intrinsic reward from interoception	Reward: $r(t) = v(t) = \varphi_v(\Pi_1 \cdot \varepsilon_1)$	δ_1 (pleasant / unpleasant surprise)
		Interoceptive prediction: $g_1(\mu_2)$	
ENVIRONMENT	Body + physical world + social world	Sensory input: $i(t) \uparrow$ / Motor output: $u(t) \downarrow$	

Signal flow: Environment $\xrightarrow{i(t)}$ Level 1 $\xrightarrow{a(t)}$ Level 2 $\xrightarrow{z(t)}$ Level 3 (ascending). Level 3 $\xrightarrow{\text{predictions}}$ Level 2 $\xrightarrow{\text{predictions}}$ Level 1 $\xrightarrow{u(t)}$ Environment (descending).

Consciousness emerges when all three RL levels operate together: the system generates its own rewards (Level 1), learns its own state representation (Level 2), and plans by recursively simulating its social world (Level 3). The "what it is like" of experience is the felt quality of this multi-level optimization — the sense of being an agent that cares about outcomes (valence), understands its situation (schemas), and knows itself through others' eyes (recursive self-model).

References for Appendix C

- Critchley, H. D., & Harrison, N. A. (2013). Visceral influences on brain and behavior. *Neuron*, 77(4), 624-638.
- Jaderberg, M., et al. (2017). Reinforcement learning with unsupervised auxiliary tasks. *ICLR 2017*.
- Ng, A. Y., Harada, D., & Russell, S. (1999). Policy invariance under reward transformations. *ICML 1999*, 278-287.
- Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1, 6.

- Schmidhuber, J. (1991). Curious model-building control systems. *IEEE International Joint Conference on Neural Networks*, 1458-1463.
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80(1), 1-27.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.
- Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *Proceedings of the 37th Allerton Conference*, 368-377.

Appendix D: Control-Theoretic Dynamics of the Strange Loop

Overview

This appendix models the Recursive Social Prediction (RSP) hierarchy as a multi-loop feedback control system. We specify reference signals, comparators, gains, and transfer functions at each level, analyze the stability conditions for the strange loop, and show how control-theoretic concepts (set-points, integral control, cascade control, stability margins) illuminate the dynamics of consciousness, emotion regulation, and psychopathology.

The formulation draws on Perceptual Control Theory (Powers, 1973, 2005), modern control theory (Åström & Murray, 2021), and the hierarchical control perspective in cognitive science (Carver & Scheier, 1998).

Notation. All symbols from Appendix A are retained. New control-theoretic notation is introduced in §D.1.

D.1 Control-Theoretic Notation

D.1.1 New Symbols

Symbol	Domain	Definition
$\mathbf{r}_l(t)$	$\mathbb{R}^{n_l}$	Reference signal (set-point) at level l
$\mathbf{p}_l(t)$	$\mathbb{R}^{n_l}$	Perceptual signal (measured output) at level l
$\mathbf{e}_l^{\text{ctrl}}(t)$	$\mathbb{R}^{n_l}$	Control error: $\mathbf{r}_l - \mathbf{p}_l$
K_l	$\mathbb{R}^{n_l \times n_l}$	Proportional gain matrix at level l
K_l^I	$\mathbb{R}^{n_l \times n_l}$	Integral gain matrix at level l
K_l^D	$\mathbb{R}^{n_l \times n_l}$	Derivative gain matrix at level l
$G_l(s)$	$\mathbb{C}^{n_l \times n_l}$	Transfer function of plant (controlled process) at level l
$C_l(s)$	$\mathbb{C}^{n_l \times n_l}$	Transfer function of controller at level l
$L_l(s)$	$\mathbb{C}^{n_l \times n_l}$	Loop transfer function: $L_l = C_l \cdot G_l$
$\lambda_{\max}(L_l)$	$\mathbb{C}$	Maximum eigenvalue of L_l (determines stability)
ϕ_m	$\mathbb{R}$	Phase margin (degrees)
g_m	$\mathbb{R}$	Gain margin (dB)
τ_l	$\mathbb{R}_{>0}$	Time constant of level l dynamics

D.1.2 RSP-to-Control Mapping

Control Concept	RSP Equivalent
Reference signal $\mathbf{r}_l$	Top-down prediction from level $l + 1$: $g_{l+1}(\boldsymbol{\mu}_{l+1})$
Perceptual signal $\mathbf{p}_l$	Bottom-up input: actual state at level l
Control error $\mathbf{e}_l^{\text{ctrl}}$	Prediction error $\boldsymbol{\varepsilon}_l$ (with sign reversed)
Controller output	Expectation update $\dot{\boldsymbol{\mu}}_l$ (perception) or action $\mathbf{u}$
Plant	Body + environment + social world
Set-point	Homeostatic target (Level 1), schema expectation (Level 2), self-concept (Level 3)
Cascade control	Three-level hierarchy with nested feedback loops

D.2 The RSP as a Cascade Control System

D.2.1 Cascade Architecture

Cascade control uses nested feedback loops: an outer (slow) loop sets the reference for a middle loop, which sets the reference for an inner (fast) loop. The RSP hierarchy is a three-level cascade:

Level	Reference	Sensor	Error	Output	Time Constant
Level 3: Outer Loop	Self-concept r_3 (who I am)	Social feedback $\rightarrow p_3$	$e_3 = r_3 - p_3$ (self-model discrepancy)	Schema set-points r_2	$\tau_3 \sim$ minutes–hours
Level 2: Middle Loop	Schema targets r_2 (expected situation)	Multimodal sensory $\rightarrow p_2$	$e_2 = r_2 - p_2$ (schema mismatch)	Interoceptive set-points r_1	$\tau_2 \sim$ seconds–minutes
Level 1: Inner Loop	Homeostatic targets r_1 (body state)	Interoceptive signals $\rightarrow p_1$	$e_1 = r_1 - p_1$ (homeostatic deviation)	Autonomic/motor action $u(t)$	$\tau_1 \sim$ milliseconds–seconds
Plant	Body + environment + social world				

Signal flow: Level 3 $\xrightarrow{r_2}$ Level 2 $\xrightarrow{r_1}$ Level 1 $\xrightarrow{u(t)}$ Plant (descending references). Plant $\xrightarrow{i(t)}$ Level 1 $\xrightarrow{p_2}$ Level 2 $\xrightarrow{p_3}$ Level 3 (ascending feedback).

D.2.2 Level-Specific Control Laws

Level 1: Homeostatic Controller (PID)

The inner loop implements proportional-integral-derivative (PID) control on body state:

$$\mathbf{u}(t) = K_1 \cdot \mathbf{e}_1^{\text{ctrl}}(t) + K_1^I \int_0^t \mathbf{e}_1^{\text{ctrl}}(\tau) d\tau + K_1^D \cdot \dot{\mathbf{e}}_1^{\text{ctrl}}(t)$$

where:

- $\mathbf{e}_1^{\text{ctrl}}(t) = \mathbf{r}_1(t) - \mathbf{i}(t)$ is the homeostatic error
- K_1 is the proportional gain (immediate response to deviation)
- K_1^I is the integral gain (eliminates steady-state error — the body doesn't tolerate persistent homeostatic deviation)
- K_1^D is the derivative gain (anticipatory response to rate of change)

The integral term is critical: it ensures zero steady-state error for homeostatic variables (temperature, blood glucose, pH). Biologically, this maps to the hypothalamic set-point mechanisms.

Level 2: Schema Controller (Proportional with Adaptive Gain)

$$\dot{\boldsymbol{\mu}}_2(t) = K_2(t) \cdot \mathbf{e}_2^{\text{ctrl}}(t) + K_2^D \cdot \dot{\mathbf{e}}_2^{\text{ctrl}}(t)$$

where:

- $\mathbf{e}_2^{\text{ctrl}}(t) = g_3(\boldsymbol{\mu}_3(t)) - \boldsymbol{\mu}_1(t)$ is the schema mismatch
- $K_2(t) = \boldsymbol{\Pi}_2(t)$ – the gain is the precision matrix, making this an **adaptive controller** whose gain changes with confidence

The attention schema s_{attn} modulates K_2 for each schema channel, implementing **gain scheduling**: different schemas receive different control authority depending on the attentional state.

Level 3: Self-Model Controller (Predictive with Recursion)

$$\dot{\boldsymbol{\mu}}_3(t) = K_3 \cdot \mathbf{e}_3^{\text{ctrl}}(t) + K_3^{\text{pred}} \cdot \hat{\mathbf{e}}_3^{\text{ctrl}}(t + 1)$$

where:

- $\mathbf{e}_3^{\text{ctrl}}(t) = \mathcal{M}^{(0)}(t) - \mathcal{M}^{(D_{\text{max}})}(t)$ is the discrepancy between direct self-model and socially-reflected self-model
- $\hat{\mathbf{e}}_3^{\text{ctrl}}(t + 1)$ is the predicted future self-model error (model predictive control)
- K_3^{pred} is the predictive gain

This is **Model Predictive Control (MPC)**: the controller uses a model of future dynamics to optimize actions over a planning horizon.

D.3 Transfer Functions and Frequency Analysis

D.3.1 Linearized Transfer Functions

Linearizing around the operating point $(\boldsymbol{\mu}_l^*, \mathbf{r}_l^*)$, each level has a transfer function in the Laplace domain ($s = j\omega$):

Level 1 Plant:

$$G_1(s) = \frac{\mathbf{P}_1(s)}{\mathbf{U}(s)} = \frac{A_1}{s + \frac{1}{\tau_1}}$$

where A_1 is the plant gain (sensitivity of interoceptive state to motor commands) and τ_1 is the body's time constant (e.g., ~2 seconds for heart rate response).

Level 1 Controller (PID):

$$C_1(s) = K_1 + \frac{K_1^I}{s} + K_1^D \cdot s$$

Level 1 Loop Transfer Function:

$$L_1(s) = C_1(s) \cdot G_1(s) = \frac{(K_1 s + K_1^I + K_1^D s^2) \cdot A_1}{s(s + \frac{1}{\tau_1})}$$

The closed-loop transfer function for each level

l

:

$$T_l(s) = \frac{L_l(s)}{1 + L_l(s)}$$

D.3.2 Timescale Separation

For the cascade to be stable, the inner loop must be faster than the outer loop. The timescale separation condition:

$$\tau_1 \ll \tau_2 \ll \tau_3$$

Typical biological values:

Level	Time Constant τ_l	Bandwidth $\omega_l = 1/\tau_l$	Biological Process
$l = 1$	$\tau_1 \approx 0.5\text{--}2 \text{ s}$	$0.5\text{--}2 \text{ Hz}$	Autonomic reflexes, heart rate, respiration
$l = 2$	$\tau_2 \approx 5\text{--}30 \text{ s}$	$0.03\text{--}0.2 \text{ Hz}$	Schema updates, emotional episodes, reorientation
$l = 3$	$\tau_3 \approx 60\text{--}3600 \text{ s}$	$0.0003\text{--}0.017 \text{ Hz}$	Self-model revision, social recalibration

The ratio

$$\tau_{l+1}/\tau_l \geq 5$$

ensures that each level's controller "sees" the inner loop as essentially instantaneous — a standard cascade design principle.

D.4 Stability Analysis

D.4.1 Local Stability: Eigenvalue Analysis

The linearized dynamics of the full three-level system:

$$\dot{\boldsymbol{\mu}} = \mathbf{A} \cdot \boldsymbol{\mu} + \mathbf{B} \cdot \mathbf{r}$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1; \boldsymbol{\mu}_2; \boldsymbol{\mu}_3]$ is the stacked state vector and $\mathbf{A}$ is the Jacobian of the coupled dynamics:

$$\mathbf{A} = \begin{pmatrix} -K_1 G'_1 & J_{12} & 0 \\ J_{21} & -K_2 G'_2 & J_{23} \\ 0 & J_{32} & -K_3 G'_3 + J_{33}^{\text{rec}} \end{pmatrix}$$

where:

- Diagonal blocks $-K_l G'_l$ are the local feedback terms (stabilizing when negative)
- Off-diagonal blocks J_{lm} capture inter-level coupling
- J_{33}^{rec} is the self-referential Jacobian from the recursive loop

Stability condition: All eigenvalues of $\mathbf{A}$ must have negative real parts:

$$\text{Re}(\lambda_i(\mathbf{A})) < 0 \quad \forall i$$

D.4.2 The Strange Loop Stability Condition

The recursive self-model at Level 3 creates a positive feedback path (self-referential loop). For stability, this positive feedback must be dominated by the negative feedback from prediction error correction:

$$\rho(J_{33}^{\text{rec}}) < \sigma_{\min}\left(K_3 \cdot \frac{\partial g_3}{\partial \mu_3^{\text{rec}}}\right)$$

where $\rho(\cdot)$ is the spectral radius and $\sigma_{\min}(\cdot)$ is the smallest singular value. In plain terms: the rate at which the self-model amplifies itself through recursive social reflection must be less than the rate at which prediction errors correct the self-model.

Convergence of the strange loop. The recursion $\mathcal{M}^{(d)} = (1 - \beta_d)\mathcal{M}^{(d-1)} + \beta_d \hat{\mathcal{M}}^{(d)}$ converges if:

$$\beta_d = \beta_0 \gamma^d \quad \text{with} \quad \beta_0 < 1, \gamma < 1$$

The total gain of the recursive path is:

$$G_{\text{rec}} = \sum_{d=1}^{D_{\max}} \beta_0 \gamma^d = \beta_0 \cdot \frac{\gamma(1 - \gamma^{D_{\max}})}{1 - \gamma}$$

For $\beta_0 = 0.3, \gamma = 0.5, D_{\max} = 5$: $G_{\text{rec}} \approx 0.29$ – well below 1, ensuring stable convergence.

D.4.3 Phase and Gain Margins

For each loop, the phase margin ϕ_m and gain margin g_m quantify robustness:

$$\phi_m = 180 + \angle L_l(j\omega_c)$$

where ω_c is the gain crossover frequency ($|L_l(j\omega_c)| = 1$).

Estimated minimum margins for stable consciousness (order-of-magnitude values based on engineering analogy with cascade control systems; empirical neural measurements would refine these):

Level	Minimum ϕ_m	Minimum g_m	Interpretation
$l = 1$	> 45	> 6 dB	Stable autonomic regulation without oscillation
$l = 2$	> 30	> 4 dB	Schema updates converge without rumination
$l = 3$	> 20	> 3 dB	Self-model stable under social perturbation

The margins decrease at higher levels because the plant becomes less predictable (social dynamics are more uncertain than autonomic dynamics), and the system tolerates more oscillation (mild self-doubt is normal; mild body temperature oscillation is not). These values follow standard cascade design practice for fast/moderate/slow loops (Skogestad & Postlethwaite, 2005) and should be interpreted as predictions to be tested against neural data, not as established empirical parameters.

D.5 Disturbance Rejection and Emotion Regulation

D.5.1 Disturbance Model

External disturbances enter at each level:

$$\mathbf{p}_l(t) = G_l(s) \cdot \mathbf{u}_l(t) + \mathbf{d}_l(t)$$

where

$$\mathbf{d}_l$$

is the disturbance:

- $\mathbf{d}_1$: unexpected interoceptive events (pain, illness, drug effects)
- $\mathbf{d}_2$: unexpected environmental changes (threats, surprises, novelty)
- $\mathbf{d}_3$: unexpected social events (rejection, praise, betrayal)

D.5.2 Sensitivity Functions

The sensitivity function $S_l(s)$ determines how well level l rejects disturbances:

$$S_l(s) = \frac{1}{1 + L_l(s)}$$

- $|S_l(j\omega)| < 1$: disturbances at frequency ω are attenuated (effective regulation)
- $|S_l(j\omega)| > 1$: disturbances at frequency ω are amplified (regulation failure)

The **complementary sensitivity** $T_l(s) = 1 - S_l(s)$ determines reference tracking. The fundamental constraint:

$$S_l(s) + T_l(s) = 1$$

This means perfect disturbance rejection and perfect reference tracking cannot coexist — there is always a trade-off. In emotional terms: you cannot simultaneously be perfectly responsive to new emotional information (low S) and perfectly stable in your self-concept (high T). This is the control-theoretic formulation of the flexibility-stability trade-off in emotion regulation.

D.5.3 Emotion Regulation Strategies as Control Reconfigurations

Different emotion regulation strategies correspond to different modifications of the control architecture:

Strategy	Control Operation	Mathematical Effect
Reappraisal	Change $\mathbf{r}_2$ (schema reference)	New set-point: $\mathbf{r}'_2 = f_{\text{reappraise}}(\mathbf{r}_2, \text{context})$
Suppression	Reduce K_1 (gain reduction)	Attenuated autonomic response: $K'_1 = \alpha K_1$, $\alpha < 1$
Attention deployment	Modify $\mathbf{\Pi}_2$ (precision/gain)	Selective gain: $K'_{2,k} \rightarrow 0$ for distressing schema k
Situation selection	Change plant input (avoidance)	Eliminate disturbance: $\mathbf{d}_l = 0$ by removing source
Acceptance	Increase S_l bandwidth	Allow disturbance passage without amplification
Rumination	Positive feedback in Level 3 loop	$G_{\text{rec}} \rightarrow 1$ (marginal stability, oscillation)

Reappraisal is the most powerful strategy because it operates at Level 2–3, changing the reference signals that the lower loops track. This aligns with empirical evidence that cognitive reappraisal is more effective than suppression for long-term emotion regulation (Gross, 2002).

D.6 Pathological Dynamics

D.6.1 Instability Modes

Control systems fail in characteristic ways. Each failure mode maps to a recognizable psychopathological pattern:

Limit cycle oscillation ($\phi_m \rightarrow 0$):

When the phase margin at Level 3 approaches zero, the self-model oscillates between states without converging:

$$\mathcal{M}(t) \approx \mathcal{M}^* + A \sin(\omega_{\text{osc}} t)$$

Clinical analog: **Bipolar disorder** — the self-model oscillates between grandiosity ($\mathcal{M}$ overestimates social standing) and depression ($\mathcal{M}$ underestimates). The oscillation frequency ω_{osc} corresponds to the cycle period (weeks to months).

Integrator windup (K_1^I accumulates during saturation):

When the Level 1 controller’s integral term accumulates error during prolonged disturbance that exceeds the actuator’s capacity:

$$\int_0^T \mathbf{e}_1^{\text{ctrl}}(\tau) d\tau \rightarrow \infty \quad \text{while} \quad \mathbf{u}_{\text{max}} \text{ is saturated}$$

Clinical analog: **Chronic stress** — the body’s homeostatic controller accumulates error (allostatic load) when the stressor exceeds the capacity for autonomic correction. When the stressor is removed, the

accumulated integral term causes overshoot (burnout, immune dysfunction).

Positive feedback runaway ($G_{\text{rec}} \geq 1$):

When the recursive self-model gain exceeds unity, the strange loop amplifies rather than converges:

$$\|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\| > \|\mathcal{M}^{(d-1)} - \mathcal{M}^{(d-2)}\| \quad \forall d$$

Clinical analog: **Paranoid ideation** — "they think I think they think..." escalates rather than settling, because each recursion amplifies the social threat signal.

Loss of reference signal ($\mathbf{r}_3 \rightarrow \mathbf{0}$):

When the Level 3 reference signal (self-concept) degrades:

$$\|\mathbf{r}_3\| \rightarrow 0 \implies \mathbf{e}_3^{\text{ctrl}} \rightarrow -\mathbf{p}_3$$

Clinical analog: **Depersonalization/derealization** — without a clear self-concept, the controller has no target; perception continues but lacks the self-referential structure that gives experience its "mine-ness."

Sensor failure ($\mathbf{p}_l \rightarrow \text{noise}$):

When the perceptual signal is dominated by noise rather than veridical information:

$$\text{SNR}_l = \frac{\|\mathbf{p}_l^{\text{signal}}\|}{\|\mathbf{p}_l^{\text{noise}}\|} \rightarrow 0$$

Clinical analog: **Psychosis** — sensory signals (Level 1–2) or social signals (Level 3) become unreliable, and the controller acts on hallucinated or delusional input.

D.6.2 Robustness and Resilience

Robustness is the system's ability to maintain performance under parameter variation (gain changes, time constant shifts). Formally:

$$\text{Robustness} = \min_l \min \left(\frac{\phi_{m,l}}{\phi_{m,l}^{\min}}, \frac{g_{m,l}}{g_{m,l}^{\min}} \right)$$

Resilience is the system's ability to return to stable operation after a large disturbance. This depends on the **region of attraction** — the set of initial conditions from which the nonlinear system converges to the desired operating point.

The RSP model predicts that resilience is enhanced by:

1. **Wide timescale separation** ($\tau_3/\tau_1 \gg 1$): prevents fast disturbances from destabilizing slow loops
2. **Moderate recursive gain** ($G_{\text{rec}} \approx 0.3$): enough self-awareness without self-amplification
3. **Flexible precision** (Π_l adaptive): attention can reallocate to compensate for disturbances
4. **Rich schema repertoire** (high $|\mathcal{C}|$): more conceptual categories = more reappraisal options = more reference signals to choose from

D.7 The Strange Loop as a Fixed-Point Iteration

The pathological dynamics of §D.6 showed what happens when the control system fails. We now provide the formal tool for understanding *why* the strange loop is normally stable — the contraction mapping theorem — and what it means precisely for the recursive self-model to converge.

D.7.1 Fixed-Point Formulation

The recursive self-model can be formulated as a fixed-point problem. Define the mapping:

$$\Psi : \mathcal{M} \mapsto (1 - \beta) \cdot \mathcal{M} + \beta \cdot \text{SocialReflect}(\mathcal{M}, \{o_j\})$$

where $\text{SocialReflect}(\mathcal{M}, \{o_j\})$ computes the average predicted self-as-seen-by-others. The fixed point $\mathcal{M}^*$ satisfies:

$$\mathcal{M}^* = \Psi(\mathcal{M}^*)$$

D.7.2 Contraction Mapping Theorem

Ψ is a contraction mapping if there exists $\rho \in [0, 1)$ such that:

$$\|\Psi(\mathcal{M}_a) - \Psi(\mathcal{M}_b)\| \leq \rho \cdot \|\mathcal{M}_a - \mathcal{M}_b\| \quad \forall \mathcal{M}_a, \mathcal{M}_b$$

The contraction coefficient:

$$\rho = (1 - \beta) + \beta \cdot L, \quad \text{where } L = \left\| \frac{1}{|\mathcal{O}|} \sum_j w_j \cdot \frac{\partial \hat{\mathcal{M}}_j}{\partial \mathcal{M}} \right\|$$

The fixed point exists and is unique when $\rho < 1$, which requires $L < 1$ — the social reflection mapping must be contractive (averaging over multiple social perspectives dampens extremes). The convergence rate is geometric:

$$\|\mathcal{M}^{(d)} - \mathcal{M}^*\| \leq \rho^d \cdot \|\mathcal{M}^{(0)} - \mathcal{M}^*\|$$

For $\rho = 0.5$ and convergence threshold $\delta_{\text{conv}} = 0.01$, convergence requires $d \geq \lceil \log(0.01) / \log(0.5) \rceil \approx 7$ iterations — consistent with the $D_{\text{max}} \approx 3\text{--}5$ observed empirically (some iterations contribute sub-threshold updates).

D.7.3 Multiple Fixed Points

In the nonlinear regime,

$$\Psi$$

may have multiple fixed points — stable self-concepts that the system can converge to depending on initial conditions. The basin of attraction for each fixed point determines which self-concept is adopted given the current social context.

Identity switching (e.g., code-switching between cultural identities) corresponds to the system transitioning between basins of attraction under social context changes. The transition occurs when a disturbance pushes $\mathcal{M}(t)$ across the basin boundary.

D.8 Observer Design and Consciousness

D.8.1 The Self as State Observer

In control theory, an **observer** (or state estimator) reconstructs the internal state of a plant from its inputs and outputs. The RSP self-model is a state observer for the agent's own cognitive-affective system:

$$\dot{\hat{\boldsymbol{\mu}}} = f(\hat{\boldsymbol{\mu}}, \mathbf{u}) + \mathbf{L} \cdot (\mathbf{y} - \hat{\mathbf{y}})$$

where:

- $\hat{\boldsymbol{\mu}}$ is the estimated internal state (self-model)
- $\mathbf{L}$ is the observer gain matrix
- $\mathbf{y} - \hat{\mathbf{y}}$ is the innovation (prediction error)

This is a Luenberger observer — the same structure as hierarchical predictive coding (Appendix A). The self-model is literally an observer of the self, in the technical control-theoretic sense.

D.8.2 Observability and Consciousness

A system is **observable** if its internal state can be reconstructed from its outputs. The observability matrix is:

$$\mathcal{O}_{\text{obs}} = \begin{pmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \end{pmatrix}$$

where $\mathbf{C}$ is the output matrix specifying which internal states are accessible to the self-model (i.e., which states the observer can "see"), and $\mathbf{A}$ is the system dynamics matrix from §D.4.1. The system is fully observable iff $\text{rank}(\mathcal{O}_{\text{obs}}) = n$ (full state dimension).

In the RSP model, observability is graded rather than binary: the proportion of internal states reconstructable by the self-model varies with the rank of $\mathbf{O}_{\text{obs}}$ relative to n . The control-theoretic proposal is that **conscious access tracks self-observability** — the states that reach awareness are precisely those reconstructable by the Level 3 observer. This is a functionalist identification: it equates a mathematical property (observability rank) with a phenomenal property (degree of self-awareness). The identification is motivated by the structural correspondence between the Luenberger observer (§D.8.1) and the hierarchical predictive coding architecture (Appendix A), but it remains a philosophical commitment, not a mathematical derivation.

Full observability = full self-awareness. Partial observability = limited self-awareness (some states are "unconscious" — active but not reconstructed by the observer).

The attention schema s_{attn} determines which states are included in the output matrix $\mathbf{C}$ — implementing the RSP model's claim that attention determines the contents of consciousness.

D.8.3 Separation Principle

The **separation principle** states that controller design and observer design can be performed independently — the optimal controller for the known-state case remains optimal when paired with an optimal observer.

Applied to RSP: the organism can separately optimize (a) how it acts given what it believes about itself and (b) how it infers its own states from sensory data. This separation is approximate in the nonlinear RSP model, but it explains why self-knowledge and self-regulation can develop somewhat independently — a person can have accurate self-knowledge but poor emotional regulation, or effective regulation strategies based on inaccurate self-models.

D.9 Summary: The Control Architecture of Consciousness

$$\text{Consciousness} = \underbrace{\text{Cascade Control}}_{\text{3-level hierarchy}} + \underbrace{\text{State Observer}}_{\text{self-model}} + \underbrace{\text{Fixed-Point Recursion}}_{\text{strange loop}}$$

The RSP model, viewed through control theory, is a cascade control system where:

1. **Level 1** implements PID homeostatic control (fast inner loop)
2. **Level 2** implements adaptive gain-scheduled schema control (middle loop)
3. **Level 3** implements model predictive control with recursive state estimation (slow outer loop)
4. **The strange loop** is a fixed-point iteration converging to a stable self-concept
5. **Consciousness** is the observability of the system to its own state observer
6. **Emotions** are the control errors at each level
7. **Attention** is gain scheduling across control channels
8. **Psychopathology** is control instability (oscillation, windup, runaway, sensor failure)

The control-theoretic perspective offers something the free energy and RL frameworks do not: a precise language for **stability, robustness, and failure modes**. It predicts not only how consciousness works when it works, but how and why it breaks down — and what parameters must be restored for recovery.

References for Appendix D

- Åström, K. J., & Murray, R. M. (2021). *Feedback Systems: An Introduction for Scientists and Engineers* (2nd ed.). Princeton University Press.
- Carver, C. S., & Scheier, M. F. (1998). *On the Self-Regulation of Behavior*. Cambridge University Press.
- Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39(3), 281-291.
- Powers, W. T. (1973). *Behavior: The Control of Perception*. Aldine.
- Powers, W. T. (2005). *Behavior: The Control of Perception* (2nd ed.). Benchmark Publications.
- Skogestad, S., & Postlethwaite, I. (2005). *Multivariable Feedback Control: Analysis and Design*. Wiley.
- Stephan, K. E., et al. (2016). Allostatic self-efficacy: A metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in Human Neuroscience*, 10, 550.

Appendix E — Unified Mathematical Framework

Overview

Appendixes A–D formalize the RSP model from four distinct mathematical traditions: hierarchical predictive coding, variational free energy, unsupervised reinforcement learning, and cascade feedback control. A natural question arises: are these four separate theories, or four descriptions of one theory?

This appendix demonstrates that they are the latter. We identify a single **master variational functional** from which all four formalisms are derivable as special cases, prove key equivalences between their central quantities, and extract emergent properties visible only from the unified perspective. The result is a mathematically unified RSP theory in three equations.

E.1 Four Windows on One Architecture

In classical mechanics, the same physical dynamics admit four equivalent mathematical formulations:

- **Newtonian mechanics** foregrounds forces and accelerations ($\mathbf{F} = m\mathbf{a}$).
- **Lagrangian mechanics** foregrounds energy differences and variational principles ($\delta \int L dt = 0$).
- **Hamiltonian mechanics** foregrounds phase space and canonical transformations ($\dot{q} = \partial H / \partial p$).
- **Hamilton-Jacobi theory** foregrounds wavefronts and optimal control ($\partial S / \partial t + H = 0$).

No physicist asks "which is the real mechanics?" The physics lives in their common mathematical core; each formulation reveals different structural features and offers different computational advantages. Physicists discovered this unity over roughly two centuries (Newton 1687, Lagrange 1788, Hamilton 1835, Jacobi 1837).

Appendixes A–D stand in an analogous — though not identical — relationship. The classical mechanics equivalences are mathematically exact; the RSP correspondences are approximate, holding under the Laplace approximation with Gaussian noise near equilibrium. The analogy is aspirational in its precision but genuine in its structure:

Formalism	Mechanical Analogue	Primary Variable	What It Reveals Best
Predictive Coding (A)	Newtonian (forces)	Prediction error ε_l	Neural message-passing architecture
Free Energy (B)	Lagrangian (variational)	Free energy F	Global optimality, why three levels
Reinforcement Learning ©	Hamiltonian (state-action)	Value V , policy π	Learning, planning, development
Control Theory (D)	Hamilton-Jacobi (optimal control)	Error $\mathbf{e}_l^{\text{ctrl}}$, gain K_l	Stability, robustness, failure modes

An important disanalogy strengthens rather than weakens the argument: the four mechanical formulations were discovered independently over centuries and their equivalence was a mathematical surprise. The RSP model’s four formalisms were developed by drawing on independently established mathematical traditions (Bayesian inference, information theory, reinforcement learning, control engineering), and the fact that they converge on the same dynamics when applied to a three-level recursive hierarchy is evidence that the architecture captures genuine structure — not that the author designed them to agree.

The strategy of this appendix is to: (i) define a unified notation, (ii) identify the master functional, (iii) prove equivalences, (iv) identify divergences, and (v) extract emergent insights.

E.2 Master Variable Correspondence

Table E.1 is the Rosetta Stone of the RSP model, mapping every core variable across all four formalisms. We introduce a **unified notation** (leftmost column) used throughout this appendix.

Table E.1a: State and Error Variables

Unified	Pred. Coding (A)	Free Energy (B)	RL ©	Control (D)	Interpretation
$\mathbf{x}_l$	$\boldsymbol{\mu}_l$ (expectations)	$\boldsymbol{\mu}_l$ (sufficient statistics)	$\mathbf{s}_l^{\text{RL}}$ (state)	$\mathbf{p}_l$ (perceptual signal)	Internal belief at level l
$\mathbf{e}_l$	$\varepsilon_l = \text{input}_l - g_l(\boldsymbol{\mu}_{l+1})$	$\boldsymbol{\varepsilon}_l$ (accuracy term)	δ_l (TD error; see §E.4)	$\mathbf{e}_l^{\text{ctrl}} = \mathbf{r}_l - \mathbf{p}_l = -\varepsilon_l$	Discrepancy driving update
$\mathbf{W}_l$	$\boldsymbol{\Pi}_l$ (precision)	$\boldsymbol{\Pi}_l$ (precision in F_l)	exploration-exploitation	K_l (gain); $K_2 = \boldsymbol{\Pi}_2$ at Level 2	Error weighting / attention
$\boldsymbol{\eta}_l$	temporal pred. error	F_{temporal} term	related to TD error across time	$\dot{\mathbf{e}}_l^{\text{ctrl}}$ (derivative term)	Temporal consistency
$\mathcal{M}$	self-model state $\mathcal{M}(t)$	$\boldsymbol{\mu}_3^{\text{rec}}$	Level 3 planning state	state being controlled at Level 3 (not the reference r_3)	Recursive identity

Table E.1b: Dynamics and Convergence Variables

Unified	Pred. Coding (A)	Free Energy (B)	RL ©	Control (D)	Interpretation
$\mathcal{J}$	$\sum_l \xi_l ^2$	$F = \sum_l F_l$	$-V$ (negative value; see §E.4)	$\sum_l \mathbf{e}_l^{\text{ctrl}} _{K_l}^2$	Objective to minimize
$g_l(\cdot)$	generative function	generative model	transition model $\mathcal{T}$	plant $G_l(s)$	How level $l + 1$ predicts l
$\dot{\mathbf{x}}_l$	$\dot{\boldsymbol{\mu}}_l$ (message passing)	$-\kappa_l \partial F / \partial \boldsymbol{\mu}_l$	TD update	$K_l \mathbf{e}_l^{\text{ctrl}}$	Dynamics of belief revision
τ_l	implicit in κ_l	$\kappa_1 \gg \kappa_2 \gg \kappa_3$	three learning phases	$\tau_1 \ll \tau_2 \ll \tau_3$	Timescale separation
$\mathbf{S}_l$	$\boldsymbol{\Sigma}_l$ (transition noise)	$\boldsymbol{\Sigma}_l$ (in F_{temporal})	—	—	Temporal noise covariance

Note on the RL column: The mapping between prediction errors (ε_l) and TD errors (δ_l) is not a direct identity but an approximate correspondence that holds at fast timescales (see Proposition E.2). The mapping between $\boldsymbol{\eta}_l$ (temporal prediction errors) and TD errors is also indirect: both capture temporal mismatch, but TD errors additionally involve the discount factor and value function.

Note on $\mathcal{M}$: In App D, the self-model $\mathcal{M}$ is the *state being controlled* at Level 3, not the reference signal r_3 . The reference r_3 represents the desired self-concept (set-point), while $\mathcal{M}^{(D_{\text{max}})}$ is the current self-model state. The control error $e_3^{\text{ctrl}} = \mathcal{M}^{(0)} - \mathcal{M}^{(D_{\text{max}})}$ measures the discrepancy between direct and socially-reflected self-model.

Convention: Throughout this appendix, we use the unified symbols $\mathbf{x}_l, \mathbf{e}_l, \mathbf{W}_l, \mathcal{J}$ and note the formalism-specific reading in parentheses where needed.

E.3 The Unified Optimization Principle

E.3.1 The Master Functional

All four appendixes minimize — explicitly or implicitly — a single variational functional:

$$|\mathcal{J}[\mathbf{x}, \mathbf{W}, \mathbf{u}] = \underbrace{\sum_{l=1}^3 \frac{1}{2} \mathbf{e}_l^T \mathbf{W}_l \mathbf{e}_l}_{\text{accuracy}} + \underbrace{\sum_{l=1}^3 \frac{1}{2} \boldsymbol{\eta}_l^T \mathbf{S}_l^{-1} \boldsymbol{\eta}_l}_{\text{temporal consistency}} - \underbrace{\frac{1}{2} \sum_{l=1}^3 \ln |\mathbf{W}_l|}_{\text{complexity}} + \underbrace{\sum_{d=1}^{D_{\max}} \gamma^d \frac{1}{2} \|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\|_{\mathbf{W}_3^{(d)}}^2}_{\text{recursive self-consistency}}| \quad (\text{E.1})$$

where:

- $\mathbf{e}_l$ = $\text{input}_l - g_l(\mathbf{x}_{l+1})$ is the prediction error at level l
- $\boldsymbol{\eta}_l = \mathbf{x}_l(t) - f_l(\mathbf{x}_l(t-1), \mathbf{x}_{l+1}(t-1))$ is the temporal prediction error
- $\mathbf{W}_l$ is the precision/weighting matrix (positive definite), corresponding to $\boldsymbol{\Pi}_l$ in Apps A–B and to K_l (at Level 2) in App D
- $\mathbf{S}_l$ is the temporal transition noise covariance ($\boldsymbol{\Sigma}_l$ in Apps A–B)
- $\mathcal{M}^{(d)}$ is the self-model at recursion depth d
- $\gamma \in (0, 1)$ is the recursion discount factor (distinct from γ_{RL} , the temporal RL discount in App C; the two are related by $\gamma_{\text{RL}} = e^{-1/\tau_3}$, see Proposition E.2)

The four terms have clear functional roles:

1. **Accuracy:** Penalizes mismatch between predictions and observations at each level.
2. **Temporal consistency:** Penalizes violations of expected temporal dynamics.
3. **Complexity:** Prevents overfitting by penalizing extreme precision (Occam’s razor).
4. **Recursive self-consistency:** Penalizes discrepancy between the direct self-model and the socially-reflected self-model, with geometrically decaying weight at each recursion depth.

E.3.2 Each Formalism as a Projection

Each appendix foregrounds different terms of

$\mathcal{J}$

and reads the minimization through a different lens:

Predictive Coding (A) focuses on the accuracy terms. The prediction errors $\mathbf{e}_l$ and their precision-weighted versions $\boldsymbol{\xi}_l = \mathbf{W}_l \mathbf{e}_l$ are the primary quantities. The dynamics are read as bidirectional message passing: bottom-up error signals and top-down predictions. What is foregrounded: the architecture of neural circuits — which neurons carry errors, which carry predictions, and how precision modulates their interaction.

Free Energy (B) identifies $\mathcal{J} = F$ directly and proves the variational bound $F \geq -\ln p(\mathbf{y})$. The decomposition $F = \text{accuracy} - \text{complexity}$ (with the temporal and recursive terms as additional

components) provides the Bayesian justification. What is foregrounded: global optimality and the information-theoretic meaning of each term.

Reinforcement Learning © identifies reward as the negative rate of change of Level 1 free energy: $r(t) \approx -\dot{F}_1(t) \cdot \Delta t$ (valence tracks how quickly interoceptive surprise is changing; see Proposition E.2 for the precise conditions). The value function is then the expected cumulative discounted reward. What is foregrounded: the sequential decision problem, the learning architecture, and the developmental ordering.

Control Theory (D) makes the identification $\mathbf{e}_l^{\text{ctrl}} = -\boldsymbol{\varepsilon}_l$ (sign convention). The gain-precision relationship is exact at Level 2 ($K_2 = \mathbf{\Pi}_2$, App D §D.2.2) and approximate at Levels 1 and 3, where the control laws have richer structure (PID at Level 1, model-predictive at Level 3). What is foregrounded: stability conditions, robustness margins, and pathological failure modes.

E.3.3 The Master Dynamics

Minimizing $\mathcal{J}$ with respect to the internal states $\mathbf{x}_l$ yields the master dynamics equation:

$$|\dot{\mathbf{x}}_l = -\kappa_l \frac{\partial \mathcal{J}}{\partial \mathbf{x}_l} = \kappa_l \left[\underbrace{\frac{\partial g_{l-1}^T}{\partial \mathbf{x}_l} \mathbf{W}_{l-1} \mathbf{e}_{l-1}}_{\text{bottom-up}} - \underbrace{\mathbf{W}_{l+1} \mathbf{e}_{l+1}}_{\text{top-down}} - \underbrace{\mathbf{S}_l^{-1} \boldsymbol{\eta}_l}_{\text{temporal}} \right] | \quad (\text{E.2})$$

This single equation simultaneously implements:

- **Message passing** (App A): The first term is the bottom-up precision-weighted prediction error from the level below (via the Jacobian of the generative function g_{l-1} that maps from level l); the second is the top-down prediction error from the level above.
- **Gradient descent on free energy** (App B): The entire expression is $-\kappa_l \partial F / \partial \boldsymbol{\mu}_l$.
- **A component of value maximization** (App C): Under the $V \approx -\mathcal{J}$ identification, this update moves the state in the direction of increasing value (see Proposition E.3 for the approximation conditions).
- **The control law** (App D): With $K_l = \kappa_l \mathbf{W}_l$, this is the controller output that drives the error toward zero.

Similarly, minimizing

$$\mathcal{J}$$

with respect to the precision matrices and the action yields:

$$|\dot{\mathbf{W}}_l = -\eta_l \frac{\partial \mathcal{J}}{\partial \mathbf{W}_l} = \eta_l \left(\frac{1}{2} \mathbf{W}_l^{-1} - \frac{1}{2} \mathbf{e}_l \mathbf{e}_l^T \right) | \quad (\text{E.3})$$

$$|\mathbf{u}^* = \arg \min_{\mathbf{u}} \mathbb{E} \left[\sum_{\tau=t}^{t+T} \mathcal{J}(\tau) \mid \mathbf{u} \right] | \quad (\text{E.4})$$

These three equations — belief update, precision update, action selection — constitute the complete RSP theory.

E.4 Equivalence Proofs

Proposition E.1: The Universal Error Signal

Statement: Under the Laplace approximation with Gaussian noise, the following quantities are equal up to sign conventions and constant factors:

(a) Precision-weighted prediction error:

$$\boldsymbol{\xi}_l = \boldsymbol{\Pi}_l \boldsymbol{\varepsilon}_l$$

(App A)

(b) Gradient of the accuracy term:

$$\partial F_l^{\text{acc}} / \partial \boldsymbol{\varepsilon}_l = \boldsymbol{\Pi}_l \boldsymbol{\varepsilon}_l$$

(App B)

© Weighted control error at Level 2:

$$K_2 \mathbf{e}_2^{\text{ctrl}} = \boldsymbol{\Pi}_2 \boldsymbol{\varepsilon}_2$$

(App D)

Proof: (a) = (b) follows from differentiating $F_l^{\text{acc}} = \frac{1}{2} \boldsymbol{\varepsilon}_l^T \boldsymbol{\Pi}_l \boldsymbol{\varepsilon}_l$ with respect to $\boldsymbol{\varepsilon}_l$. For (a) = © at Level 2: the control error is $\mathbf{e}_2^{\text{ctrl}} = g_3(\boldsymbol{\mu}_3) - \boldsymbol{\mu}_1 = -\boldsymbol{\varepsilon}_2$, and $K_2 = \boldsymbol{\Pi}_2$ (App D, §D.2.2), giving $K_2 \mathbf{e}_2^{\text{ctrl}} = -\boldsymbol{\xi}_2$. The sign difference reflects the convention that prediction errors point from prediction to observation while control errors point from set-point to measurement.

Remark: At Levels 1 and 3, the control gains have richer structure (PID at Level 1, model-predictive at Level 3) and the exact identity $K_l = \boldsymbol{\Pi}_l$ does not hold. However, the proportional component of each control law is precision-weighted, so the correspondence holds approximately for the dominant term. $\square$

Proposition E.2: Free Energy–Value Duality

Statement: Under the following conditions — (i) the Laplace approximation holds, (ii) valence is defined as the precision-weighted interoceptive prediction error $v(t) = \varphi_v(\boldsymbol{\Pi}_1 \cdot \boldsymbol{\varepsilon}_1)$ (App C, §C.2.1), and (iii) the RL discount is related to the slowest timescale by $\gamma_{\text{RL}} = e^{-1/\tau_3}$ — the following approximate duality holds:

$$V(\mathbf{z}) \approx -\mathbb{E} \left[\sum_{\tau=0}^{\infty} \gamma_{\text{RL}}^{\tau} F_1(t + \tau) \mid \mathbf{z}(t) = \mathbf{z} \right] + \text{const}$$

where F_1 is the Level 1 free energy (not total F).

Proof sketch: The value function is the expected cumulative discounted reward: $V = \mathbb{E}[\sum \gamma_{\text{RL}}^\tau r(t + \tau)]$. Reward is a monotonic function of Level 1 prediction error: $r(t) = v(t) = \varphi_v(\mathbf{\Pi}_1 \cdot \boldsymbol{\varepsilon}_1)$, which under the Laplace approximation is approximately $-F_1(t)$ up to a monotonic transformation. Therefore $V \approx -\mathbb{E}[\sum \gamma_{\text{RL}}^\tau F_1(t + \tau)] + \text{const.}$

The expected free energy $G(\pi)$ from App B §B.7 involves the *total* free energy F over the planning horizon, not just F_1 . The duality $V \approx -G$ therefore holds only approximately: the value function captures primarily Level 1 contributions (interoceptive reward), while the expected free energy additionally penalizes Levels 2 and 3. The two are equivalent when Level 1 dominates the total free energy, which is the regime where the organism is primarily motivated by homeostatic concerns. At longer timescales, where Levels 2 and 3 contribute significantly, the duality is an approximation. $\square$

Corollary: The TD error $\delta_l(t) = r(t + 1) + \gamma V_l(\mathbf{x}_l(t + 1)) - V_l(\mathbf{x}_l(t))$ (Sutton & Barto, 2018) is related to the change in free energy: $\delta_l(t) \approx -\Delta F_l(t)$ at fast timescales. A positive TD error ("better than expected") corresponds to decreasing free energy ("surprise is reducing").

Proposition E.3: The Master Dynamics Unifies All Update Rules

Statement: The master dynamics $\dot{\mathbf{x}}_l = -\kappa_l \partial \mathcal{J} / \partial \mathbf{x}_l$ simultaneously implements:

- (a) Bidirectional message passing with precision-weighted errors (App A, §A.5.1–A.5.3)
- (b) Gradient descent on the variational free energy bound (App B, §B.4.1)
- © An approximate value-gradient update (App C, §C.6)
- (d) The closed-loop control law with feedforward and feedback (App D, §D.2.2)

Proof: For (a) and (b), the equivalence is immediate: expanding $\partial \mathcal{J} / \partial \mathbf{x}_l$ produces the three-term expression (bottom-up error, top-down error, temporal error) that defines the message-passing scheme in App A, and $\mathcal{J} = F$ by construction in App B.

For (d), under the sign convention $\mathbf{e}_l^{\text{ctrl}} = -\boldsymbol{\varepsilon}_l$ and identification $K_l = \kappa_l \mathbf{W}_l$, the master dynamics reads $\dot{\mathbf{x}}_l = K_l \mathbf{e}_l^{\text{ctrl}} + \text{feedforward terms}$, which is the control law.

For ©, the connection is approximate rather than exact. The master dynamics performs gradient descent on $\mathcal{J}$, which under the duality $V \approx -\mathcal{J}$ (Proposition E.2) is gradient *ascent* on the value function. This is analogous to but not identical with TD learning: both converge to the same fixed point (the optimal value function satisfying the Bellman equation), but the gradient descent dynamics and the TD update rule differ in their transient trajectories. The correspondence is closest at fast timescales and near equilibrium, where the Bellman equation's fixed-point condition $\partial V / \partial \mathbf{x}_l = 0$ coincides with the master dynamics' equilibrium $\partial \mathcal{J} / \partial \mathbf{x}_l = 0$. $\square$

Proposition E.4: Universal Strange Loop Convergence

Statement: The recursive self-modeling loop converges in all four formalisms under a common condition. Define the contraction coefficient:

$$|\rho = (1 - \beta_0) + \beta_0 \cdot L, \quad \text{where } L = \left\| \frac{\partial \text{SocialReflect}}{\partial \mathcal{M}} \right\| | \quad (\text{E.5})$$

where $\beta_0 \in (0, 1)$ is the base mixing parameter and L is the Lipschitz constant of the social reflection mapping. Contraction requires $L < 1$ – the social reflection must be a smoothing operation (averaging over multiple social perspectives dampens extremes rather than amplifying them). When $\rho < 1$, the Banach fixed-point theorem guarantees convergence:

(a) **Predictive coding (A):** The recursion terminates: $\|\mathcal{M}^{(d)} - \mathcal{M}^*\| \leq \rho^d \|\mathcal{M}^{(0)} - \mathcal{M}^*\|$ (this bound follows from the contraction mapping; App A §A.4.4 states the termination condition $\|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\| < \delta_{\text{conv}}$, which is implied by the contraction)

(b) **Free energy (B):** The recursive free energy contributions decay geometrically: $F_3^{(d)}$ terms are weighted by γ^d with $\gamma < 1$, ensuring the sum F_3^{rec} converges (App B, §B.3.3)

© **RL** ©: The social reward shaping diminishes: $|r_{\text{social}}^{(d)}| \propto \gamma^d$ (App C, §C.4.3)

(d) **Control (D):** The recursive gain in the stability matrix is bounded: $\|J_{33}^{\text{rec}}\| \leq \rho < 1$ (App D, §D.4.2)

Proof sketch: The contraction mapping $\Psi_d : \mathcal{M} \mapsto (1 - \beta_d)\mathcal{M} + \beta_d \cdot \text{SocialReflect}(\mathcal{M})$ with depth-dependent mixing $\beta_d = \beta_0 \gamma^d$ is the common mathematical object across all four formalisms. The contraction coefficient at each depth is $\rho_d = (1 - \beta_d) + \beta_d L = 1 - \beta_d(1 - L)$. Since $L < 1$ and $\beta_d > 0$, we have $\rho_d < 1$ at every depth. The tightest contraction (smallest ρ_d) occurs at $d = 0$ where β_d is largest: $\rho_0 = (1 - \beta_0) + \beta_0 L$. As depth increases, $\beta_d \rightarrow 0$ and $\rho_d \rightarrow 1$, so deeper recursions are less contractive but the geometric discount γ^d ensures their contributions remain bounded. The Banach theorem guarantees convergence when $\rho_0 < 1$, which holds whenever $L < 1$. $\square$

E.5 Where the Formalisms Diverge

The cross-formalism equivalences established in §E.4 hold under specific idealizing conditions: the Laplace approximation, Gaussian noise, near-equilibrium dynamics, and the assumption that Level 1 free energy dominates total free energy (for the RL duality). The equivalences are approximate rather than exact – analogous to geometric optics and wave optics agreeing in the short-wavelength limit, not to the exact mathematical identities between Newtonian and Lagrangian mechanics. We use "equivalence" in the sense of convergence to the same fixed points and qualitative agreement on dynamics, not identity of transient trajectories or out-of-equilibrium behavior.

The formalisms genuinely diverge in several domains.

We transition from the formal unity of the master functional to the practical diversity of its four projections. Each formalism offers unique predictive and computational advantages precisely because it foregrounds different aspects of the shared structure.

E.5.1 Divergent Predictions

Domain	Pred. Coding (A)	Free Energy (B)	RL ©	Control (D)
Neural circuits	Error vs. representation neurons anatomically separable	Agnostic about circuits	Dopaminergic TD error signals	Gain-modulated feedback loops
Psychopathology	Aberrant precision	Excessive complexity cost	Reward dysfunction	Instability (oscillation, windup)
Development	Hierarchical model construction	Sequential free energy reduction	Reward $\rightarrow$ state $\rightarrow$ model ordering	Inner loops stabilize first
Measurement	fMRI prediction error signals	Surprise/information measures	Behavioral choice data	Frequency-domain stability

E.5.2 Non-Equivalent Regimes

1. **Far from equilibrium:** The Laplace approximation fails when prediction errors are large. Free energy and predictive coding presuppose near-equilibrium; RL handles large reward fluctuations more naturally; control theory handles nonlinear saturation and limit cycles.
2. **Discrete action spaces:** RL naturally handles discrete choices; control theory handles continuous regulation; predictive coding and free energy are agnostic but lean continuous.
3. **Model learning vs. model use:** RL cleanly separates learning (updating $\mathcal{T}, \mathcal{R}$) from planning (using them to select π). The other formalisms blur this — perception and learning are both gradient descent on $\mathcal{J}$.
4. **Stability analysis:** Only control theory (D) provides phase margins, gain margins, and sensitivity functions — tools for predicting exactly how and when the system will fail. These have no natural analog in the other formalisms.

E.5.3 Complementary Advantages

Problem	Best Formalism	Why
Neural circuit modeling	A	Message passing maps to cortical layers
Proving architectural optimality	B	Variational bound guarantees
Designing learning experiments	C	TD errors predict dopamine signals
Predicting clinical failure	D	Stability analysis classifies pathology
Cross-level interactions	E (Unified)	Captures all interactions simultaneously

E.6 Emergent Properties of the Unified Framework

The preceding sections established the formal unity of the four formalisms. We now turn to properties that emerge *only* from the unified perspective — structural features invisible to any single formalism.

E.6.1 Coordinate Covariance

The RSP model possesses a form of **coordinate covariance**: the accuracy term in $\mathcal{J}$ is invariant under smooth reparametrizations of the internal states. If $\mathbf{x}_l \rightarrow h_l(\mathbf{x}_l)$ for any smooth invertible h_l , and the precision matrices transform as:

$$g_l \rightarrow g_l \circ h_{l+1}^{-1}, \quad \mathbf{W}_l \rightarrow (\mathbf{J}_l^{-1})^T \mathbf{W}_l \mathbf{J}_l^{-1}$$

where $\mathbf{J}_l = \partial h_l / \partial \mathbf{x}_l$ is the Jacobian, then the accuracy terms $\frac{1}{2} \mathbf{e}_l^T \mathbf{W}_l \mathbf{e}_l$ are invariant. The complexity terms $-\frac{1}{2} \ln |\mathbf{W}_l|$ acquire an additional $\ln |\mathbf{J}_l|$ factor, so the full $\mathcal{J}$ is not strictly invariant — the invariance is partial, holding exactly for the accuracy and temporal terms but only approximately for the full functional when complexity terms are small.

The philosophical significance remains: the dominant terms in $\mathcal{J}$ (those that drive the dynamics at fast timescales) are invariant under reparametrization. What matters for the dynamics of consciousness is the relational structure — prediction errors and their weightings — not the specific neural coding scheme. This is the mathematical expression of the paper’s claim that consciousness is constituted by the *process* of meaningful integration, not by any specific neural code. It is a substantive philosophical position (akin to structural realism about consciousness), not merely a mathematical corollary.

E.6.2 Bounded Total Error

An upper bound constrains the total weighted error across the hierarchy:

$$\left| \sum_{l=1}^3 \text{tr}(\mathbf{W}_l \mathbf{e}_l \mathbf{e}_l^T) \right| \leq 2\mathcal{J} + \sum_{l=1}^3 \ln |\mathbf{W}_l| \quad (\text{E.6})$$

This follows from the non-negativity of the temporal and recursive terms in $\mathcal{J}$: dropping them from the right-hand side can only decrease it. The inequality means that reducing error at one level either reduces the total functional $\mathcal{J}$ (global improvement) or redistributes error to another level (trade-off). Precision allocation ($\mathbf{W}_l$) controls this redistribution. This explains the limited capacity of conscious experience: attentional precision is a finite resource, and allocating it to one level reduces availability at others.

E.6.3 Three-Level Convergence Argument

Claim: For a self-modeling agent that (i) generates intrinsic reward from interoception, (ii) maintains stable self-representation under social perturbation, and (iii) plans by simulating others' responses, three hierarchical levels with timescale separation constitute the natural minimum architecture.

We state this as a convergence argument rather than a formal impossibility theorem: while we cannot rule out that some entirely different architecture might achieve the same functional properties with fewer levels, all four formalisms independently reach the conclusion that three levels are needed *given the RSP model's specific assumptions*:

- **Free energy (B):** A one-level model cannot simultaneously minimize interoceptive surprise (F_1) and self-model surprise (F_3); Level 2 is required as a mediating representation that provides the structured state on which Level 3 operates (§B.5.2).
- **RL ©:** Reward generation (Level 1) must precede state representation learning (Level 2), which must precede model-based planning (Level 3). This is the developmental ordering constraint — you cannot learn a planning model before you have a state representation, and you cannot learn a state representation before you have a reward signal (§C.5.1).
- **Control (D):** Cascade stability requires inner loops to settle before outer loops. The timescale ratio $\tau_{l+1}/\tau_l \geq 5$ is a standard cascade design constraint; violating it produces coupled oscillations (§D.3.2).
- **Predictive coding (A):** Three levels of the generative model are the minimum to produce simultaneous prediction errors at interoceptive, schematic, and self-referential scales (§A.3).

The convergence of four independent arguments — each from a different mathematical tradition, each reaching the same structural conclusion — is stronger evidence than any one argument alone.

E.6.4 Precision as Universal Currency

Across all four formalisms, precision ($\mathbf{W}_l / \mathbf{\Pi}_l / K_l$) serves as the universal mediating variable:

Formalism	Role of Precision
Predictive Coding (A)	Weights prediction errors in message passing
Free Energy (B)	Balances accuracy vs. complexity; prevents overfitting
RL ©	Controls exploration (low $\mathbf{W}$) vs. exploitation (high $\mathbf{W}$)
Control (D)	Sets controller gain; determines stability margins

All four roles are the same mathematical operation: multiplicative scaling of error signals. This means **attention** — the mechanism that modulates precision — is the single process that simultaneously determines what reaches consciousness (A), what is learned (B), what is valued ©, and what is stable (D). No single formalism reveals this quadruple role; it emerges only from the unified perspective.

E.6.5 Emotion as Multi-Formalism Error

The unified framework provides the deepest account of what emotions are. An emotion at time

$$t$$

is simultaneously:

$$\text{Emotion}(t) = \left\langle \varepsilon_l, \frac{\partial F}{\partial \mu_l}, \delta_l, \mathbf{e}_l^{\text{ctrl}} \right\rangle$$

These four aspects correspond to:

- What is **surprising** (prediction error, App A)
- What **needs to change** (free energy gradient, App B)
- How things are **going** relative to expectations (TD error, App C)
- How **far we are** from our goals (control error, App D)

A single-formalism account of emotion captures only one aspect. The unified framework captures all four simultaneously, explaining why emotions are at once informational (surprising), motivational (demanding action), evaluative (good/bad), and regulatory (error-correcting). This four-fold characterization is a genuinely novel contribution that no single formalism produces alone.

E.7 Unified Falsification Conditions

E.7.1 Formalism-Independent Predictions

These predictions follow from the unified framework regardless of which formalism is used:

1. **Timescale separation:** The three levels must operate at distinct timescales with $\tau_{l+1}/\tau_l \geq 5$. Neural oscillations supporting consciousness should cluster into three frequency bands. Testable via EEG/MEG spectral analysis during tasks engaging different RSP levels.
2. **Precision necessity:** Disrupting precision modulation ($\mathbf{W}_l \rightarrow \mathbf{I}$) should degrade conscious experience without eliminating neural activity. Testable via pharmacological manipulation of neuromodulators implementing precision (acetylcholine, norepinephrine) and measuring subjective report alongside neural activity.
3. **Contraction bound:** The strange loop contraction coefficient $\rho < 1$ is required for stable consciousness. Clinical populations with self-model instability (e.g., borderline personality disorder, identity disturbance) should show $\rho \rightarrow 1$, operationalized as slower convergence in recursive self-

evaluation tasks. Psychotic states with paranoid ideation should show $\rho > 1$, operationalized as escalating self-referential rumination.

4. **Three levels necessary and sufficient:** Adding a fourth level should not improve model fit for consciousness-related data; removing any one level should produce a specific, predictable pattern of degradation (removal of Level 1: loss of valence; removal of Level 2: loss of structured experience; removal of Level 3: loss of self-awareness).

E.7.2 Formalism-Specific Falsification

Each formalism makes unique predictions that, if falsified, would undermine that representation without necessarily falsifying the unified framework:

Formalism	Unique Prediction	Falsified By
Pred. Coding (A)	Error and representation neurons are anatomically separable	All cortical neurons carry mixed signals
Free Energy (B)	F decreases monotonically during perceptual convergence	Systematic F increases during normal perception
RL ©	TD errors correlate with interoceptive prediction errors	Dopaminergic signals purely externally driven
Control (D)	Phase margins decrease from Level 1 to Level 3	Level 3 more robustly stable than Level 1

E.7.3 Falsifying the Unified Framework

The entire framework would be falsified by demonstrating:

1. **Consciousness without interoception** — a being with full self-awareness but no interoceptive monitoring (contradicts three-level architecture: no Level 1). Note: patients with congenital insensitivity to pain (CIPA) have *reduced* but not *absent* interoception — they retain temperature sensitivity, hunger, and proprioception. Full absence of interoception has not been documented in any conscious being.
2. **Consciousness without timescale separation** — a single-timescale system exhibiting all properties of conscious experience.
3. **Consciousness without self-modeling** — full phenomenal consciousness with no capacity for recursive self-representation (contradicts Level 3 as constitutive).
4. **Precision independent of conscious content** — attention and consciousness fully dissociable.

E.7.4 The AI Consciousness Question

RSP implies that an artificial system implementing the full three-level architecture — including genuine interoceptive monitoring (Level 1), schema-organized experience (Level 2), and recursive social prediction with depth

$$D_{\max} \geq 3$$

(Level 3b) — would be conscious. Current large language models do not satisfy these conditions: they lack biological core affect (no interoceptive monitoring producing genuine valence), lack embodied schemas (no body schema, no spatial schema grounded in sensorimotor interaction), and perform self-reference without recursive self-modeling (no contraction mapping converging to a stable self-model through social prediction).

However, an artificial system with learned valence functions, embodied interaction, and genuine recursive social prediction would, under RSP, be a candidate for consciousness. This is a prediction of the theory, not a dismissal. The question of whether such a system would be conscious cannot be settled a priori — it depends on whether the constitutive premise (§E.10) is correct. What RSP provides is a principled set of architectural conditions that any candidate conscious system must satisfy, rather than an unprincipled appeal to biological chauvinism or substrate independence. The falsification conditions in §E.7.1–E.7.3 apply equally to biological and artificial systems: if a system satisfies the formal conditions and is not conscious, or if a system is conscious without satisfying them, RSP is falsified.

E.8 Summary: The Mathematical Core of Consciousness

The RSP model reduces to three equations operating over a three-level recursive hierarchy:

$$\begin{array}{ll} \dot{\mathbf{x}}_l = -\kappa_l \frac{\partial \mathcal{J}}{\partial \mathbf{x}_l} & \text{(belief update: perception and learning)} \\ \dot{\mathbf{W}}_l = -\eta_l \frac{\partial \mathcal{J}}{\partial \mathbf{W}_l} & \text{(precision update: attention)} \\ \mathbf{u}^* = \arg \min_{\mathbf{u}} \mathbb{E}[\mathcal{J}] & \text{(action selection: behavior)} \end{array}$$

with the master functional:

$$\mathcal{J} = \sum_{l=1}^3 \left[\frac{1}{2} \mathbf{e}_l^T \mathbf{W}_l \mathbf{e}_l + \frac{1}{2} \boldsymbol{\eta}_l^T \mathbf{S}_l^{-1} \boldsymbol{\eta}_l - \frac{1}{2} \ln |\mathbf{W}_l| \right] + \sum_{d=1}^{D_{\max}} \gamma^d \frac{1}{2} \|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\|_{\mathbf{W}_3^{(d)}}^2$$

These equations simultaneously describe message passing (read as predictive coding), variational inference (read as free energy minimization), reward optimization (read as reinforcement learning), and error regulation (read as feedback control). The four readings are not four theories but four windows onto one mathematical structure — just as Newton, Lagrange, Hamilton, and Hamilton-Jacobi offer four windows onto classical mechanics.

What the unified framework reveals beyond any single formalism: consciousness has a **representational invariance** (the dominant dynamics are independent of the internal coding scheme), a **bounded error** property (attentional precision is finite and must be allocated), a **three-level convergence** (four independent formalisms agree that three levels with timescale separation are the natural minimum), a

universal currency (precision mediates everything), and a **four-fold characterization of emotion** (simultaneously surprise, gradient, evaluation, and error).

What is gained by unification, beyond elegance? Three things. First, it prevents false dichotomies: consciousness is not "computation *or* control" — it is a single process that admits both descriptions. Second, it generates novel predictions that no single formalism produces: the representational invariance, the error bound, and the four-fold emotion structure are emergent from the unified perspective. Third, it provides a single set of falsification conditions (§E.7) that could refute the entire framework at once, rather than requiring separate tests for each formalism.

The physics of consciousness, like the physics of motion, admits multiple equivalent mathematical descriptions. Consciousness lives not in any one description but in their common core — the self-referential, self-modeling, socially-calibrated optimization process that all four descriptions illuminate from their distinct vantage points.

E.9 Algorithmic Specification of the Unified Model

The preceding sections defined the master functional

$$\mathcal{J}$$

(§E.3), proved cross-formalism equivalences (§E.4–E.5), and identified emergent properties (§E.6). This section provides the procedural complement: a step-by-step algorithm that implements the unified RSP dynamics. Each step corresponds to a specific operation in the master functional framework; cross-references connect the pseudocode to the formal equations in Appendixes A–E.

All notation follows §A.1.1. Subroutines that remain unspecified (EstimateBeliefs, EstimateGoals, Simulate) represent active research areas in computational theory of mind; their forms are not constrained by the RSP framework.

E.9.1 Algorithm 1: Unified RSP Loop

Input: Sensory input $\mathbf{x}(t)$, interoceptive input $\mathbf{i}(t)$, social context $\mathcal{O} = \{o_1, \dots, o_K\}$

Output: Phenomenal state $\Phi(t) \in \mathbb{R}^{2n_e+n_m}$, updated self-model $\mathcal{M}(t) \in \mathbb{R}^{n_m}$

Level 1: Core Affect — gradient descent on F_1 (§A.3.1, §B.3.1)

1. $\mathbf{a}(t) \leftarrow \phi_{\text{aff}}(\mathbf{\Pi}_1 \boldsymbol{\epsilon}_1(t))$ — precision-weighted interoceptive prediction error mapped to valence + arousal

Level 2: Schema Integration — gradient descent on F_2 (§A.3.2, §B.3.2)

2. **For** each schema $s_k \in \mathcal{S} = \{s_{\text{body}}, s_{\text{spatial}}, s_{\text{aff-hom}}, s_{\text{attn}}, s_{\text{self}}\}$:
 - $s_k(t) \leftarrow f_k^{\text{schema}}(s_k(t-1), \mathbf{x}(t), \mathbf{a}(t), g_{2,k}(\boldsymbol{\mu}_3(t-1)))$ — four-argument update per §A.3.2
3. $\mathbf{z}(t) \leftarrow \mathbf{W}_{\text{int}} \cdot \text{concat}(\mathbf{a}(t), \{s_k(t)\})$ — structured affect via learned integration (§A.3.2)

4. Update precisions: $\mathbf{\Pi}_l \leftarrow \mathbf{\Pi}_l + \eta_l(\frac{1}{2}\mathbf{\Pi}_l^{-1} - \frac{1}{2}\boldsymbol{\varepsilon}_l\boldsymbol{\varepsilon}_l^T)$ for $l = 1, 2$ – attention allocation (§A.5.4, §E.3.3)

Level 3a: Conceptual Categorization – Bayesian model selection on F_3 (§A.3.3, §B.6)

5. **If** $|\mathcal{C}| > 0$: $c^* \leftarrow \arg \max_{c \in \mathcal{C}} P(c \mid \mathbf{z}(t), \text{context})$ – predictive categorization via $\arg \min_c \mathbb{E}[F_c]$
6. **If** $|\mathcal{C}| > 0$: $\mathbf{e}(t) \leftarrow \phi_{\text{exp}}(\mathbf{z}(t), \mathbf{h}_{c^*})$ – constituted experience (§A.3.3); **Else**: $\mathbf{e}(t) \leftarrow \mathbf{z}(t)$

Level 3b: Recursive Social Prediction – contraction mapping on $\mathcal{M}^*$ (§A.4.3, §D.7, §E.4)

7. $\mathcal{M}(t) \leftarrow \tanh(\mathbf{W}_m \cdot \text{concat}(\mathcal{M}(t-1), \mathbf{e}(t)) + \mathbf{b}_m)$ – self-model update (§A.4.1)
8. **For** $d = 1$ to D_{max} :
 - **For** each other-model $o_j \in \mathcal{O}$:
 - $\hat{\mathcal{M}}_j^{(d)} \leftarrow \text{PredictSelfAsSeenBy}(\mathcal{M}(t), o_j)$ – "How does o_j model me?"
 - $\hat{\mathcal{M}}^{(d)} \leftarrow \sum_{j=1}^K w_j \hat{\mathcal{M}}_j^{(d)}$, where $w_j \in [0, 1], \sum_j w_j = 1$ – salience-weighted aggregation (§A.4.3)
 - $\mathcal{M}(t) \leftarrow (1 - \beta_d)\mathcal{M}(t) + \beta_d \hat{\mathcal{M}}^{(d)}$, where $\beta_d = \beta_0 \gamma^d$ – strange loop (contraction mapping Ψ , §E.4)
 - **If** $\|\mathcal{M}(t) - \mathcal{M}_{\text{prev}}\| < \delta_{\text{conv}}$: **break** – early termination on convergence (§A.4.4)

Temporal Prediction – expected free energy (§B.5, §C.3)

9. $\hat{\mathbf{e}}(t+1) \leftarrow g_3(\mathcal{M}(t))$ – predict next experience via generative model
10. $\boldsymbol{\eta}(t) \leftarrow \mathbf{e}(t) - \hat{\mathbf{e}}_{\text{prev}}(t)$ – temporal prediction error (§A.2.3)
11. $\mathcal{M}(t) \leftarrow \mathcal{M}(t) - \kappa_3 \frac{\partial g_3^T}{\partial \mathcal{M}} \mathbf{S}_3^{-1} \boldsymbol{\eta}(t)$ – gradient correction projected from experience space to self-model space via Jacobian (§E.3.3)

Phenomenal State Assembly

12. $\Phi(t) \leftarrow (\mathbf{e}(t), \mathcal{M}(t), \hat{\mathbf{e}}(t+1)) \in \mathbb{R}^{n_e} \times \mathbb{R}^{n_m} \times \mathbb{R}^{n_e}$ – consciousness $\equiv$ this integrated state
13. **Return** $\Phi(t), \mathcal{M}(t)$

E.9.2 Subroutine: PredictSelfAsSeenBy

Input: Self-model $\mathcal{M} \in \mathbb{R}^{n_m}$, other-model o_j

Output: Predicted self-as-seen-by-other $\hat{\mathcal{M}}_j \in \mathbb{R}^{n_m}$

1. $\hat{o}_j.\text{beliefs} \leftarrow \text{EstimateBeliefs}(o_j)$ – infer o_j 's knowledge state
2. $\hat{o}_j.\text{goals} \leftarrow \text{EstimateGoals}(o_j)$ – infer o_j 's motivational state
3. $\hat{\mathcal{M}}_j \leftarrow \text{Simulate}(\mathcal{M}, \hat{o}_j.\text{beliefs}, \hat{o}_j.\text{goals})$ – forward model of o_j 's inference about me
4. **Return** $\hat{\mathcal{M}}_j$

E.9.3 Phylogenetic Parameterization

The algorithm's behavior degrades gracefully as parameters are reduced, producing the consciousness gradient predicted by the RSP model. When $|\mathcal{C}| = 0$, steps 5-6 are bypassed (no conceptual categorization). When $D_{\max} = 0$, step 8 executes zero iterations (no recursive social prediction). The complexity analysis below treats per-step subroutine costs (matrix multiplications, nonlinear activations) as constants in the architectural parameters.

Organism	$ \mathcal{S} $	$D_{\max}$	$ \mathcal{C} $	Consciousness Type	----- :- :- :- -----	
Invertebrates	0-1	0	0	Proto-affect / valenced response		Fish, amphibians, reptiles 2-3 0 0
Valenced affect + rudimentary schemas				Non-social mammals	3-4	0 0 Structured affect
Social mammals	4-5	1	0	Self-aware affect, basic social prediction		Corvids 3-4 0-1 0 Structured
affect + proto-social (convergent)				Great apes	5	1-2 Low Recursive self-modeling
Humans	5			3+	High	Full phenomenal consciousness

Total per-step architectural complexity: $O(|\mathcal{C}| + D_{\max} \cdot |\mathcal{O}|)$. The self-sustaining nature of the loop (each $\Phi(t)$ generates predictions feeding $\Phi(t + 1)$) creates the continuous stream of consciousness.

E.10 The Formal-to-Phenomenal Bridge

The master functional, the equivalence proofs, and the algorithmic specification describe a dynamical system. They do not, by themselves, establish that this system is phenomenally conscious. An additional premise is required.

E.10.1 The Constitutive Premise

RSP's central philosophical claim is that phenomenal consciousness is not a further property *added to* the right functional organization but *is* what that organization is, experienced from the perspective of the system performing it. This is a form of constitutive functionalism (Shoemaker, 2007): the organizational description — three-level recursive self-modeling with timescale separation, precision-weighted error minimization, and convergent strange-loop dynamics — is not merely correlated with consciousness but constitutive of it. On this view, asking "why does this particular dynamical system give rise to consciousness?" is akin to asking "why does this particular molecular configuration give rise to liquidity?" The answer in both cases is that the macroscopic property just *is* the microscopic organization, described at a different level of abstraction.

E.10.2 Distinguishing RSP from Adjacent Positions

This constitutive claim differs from several nearby positions:

(a) **Type-identity theory** identifies consciousness with specific neural types (e.g., particular cell populations, particular neurotransmitters). RSP identifies consciousness with a functional organization that could, in principle, be realized in different substrates — what matters is the three-level recursive architecture, not the biological material implementing it.

(b) **Epiphenomenalism** makes consciousness causally inert — a byproduct of physical processes with no causal role. RSP holds that the recursive self-model is causally indispensable: it shapes action selection (step 12 of Algorithm 1), modulates precision allocation, and drives the social prediction that calibrates it. Remove the self-model and the system’s behavior changes.

© **Mysterianism** (McGinn, 1989) claims the relationship between physical processes and consciousness is unknowable in principle. RSP claims the relationship is constitutive and knowable, though irreducible to lower-level physical description alone. The irreducibility is organizational, not metaphysical: the dynamics of $\mathcal{J}$ cannot be fully characterized without reference to the three-level recursive structure, but nothing about that structure is mysterious or beyond empirical investigation.

E.10.3 Falsifiability and the Constitutive Claim

If the constitutive premise is wrong — if a system could satisfy all the formal conditions specified in §E.1–E.9 without being conscious — then RSP would be falsified as a theory of consciousness, though it would survive as a theory of recursive self-modeling. Conversely, if a system were shown to be conscious without satisfying these conditions, the formal framework would be falsified as a set of necessary conditions. The falsification conditions in §E.7 are designed to test precisely this: they specify what the theory predicts about systems that satisfy (or fail to satisfy) the formal requirements.

E.10.4 What the Mathematics Does and Does Not Provide

Whether the constitutive premise is correct cannot be settled by mathematics alone. What the formal treatment provides is precision about *what* is claimed to be conscious — a system satisfying $\mathcal{J}$, with convergent recursive self-modeling at depth $D_{\max} \geq 3$, operating at three timescales with $\tau_{l+1}/\tau_l \geq 5$ — and *how* to test whether such systems are indeed conscious (§E.7). The mathematics converts a vague philosophical claim ("consciousness arises from self-modeling") into a precise empirical one ("consciousness is constituted by a system satisfying these specific formal conditions, and here is how to falsify that claim"). The gap between formal description and phenomenal reality is not closed by adding more equations. It is closed — if it is closed — by the constitutive premise that the formal description and the phenomenal reality are two descriptions of the same thing.

References for Appendix E

- Goldstein, H., Poole, C., & Safko, J. (2001). *Classical Mechanics* (3rd ed.). Addison-Wesley. [Classical mechanics analogy: four formulations of one physics]
- Kappen, H. J. (2005). Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics*, P11011. [Free energy and optimal control connection]
- Levine, S. (2018). Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv:1805.00909*. [RL-as-inference unification]
- McGinn, C. (1989). Can we solve the mind-body problem? *Mind*, 98(391), 349–366. [Mysterianism]

- Rawlik, K., Toussaint, M., & Vijayakumar, S. (2013). On stochastic optimal control and reinforcement learning by approximate inference. *Robotics: Science and Systems IX*. [Stochastic control and RL equivalence]
- Shoemaker, S. (2007). *Physical Realization*. Oxford University Press. [Constitutive functionalism]
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. [TD error definition and Bellman optimality]
- Todorov, E. (2008). General duality between optimal control and estimation. *47th IEEE Conference on Decision and Control*, 4286–4292. [Control-estimation duality]

Appendix F — Equation Reference Sheet

This appendix collects the defining equations of the RSP framework from Appendixes A through E. Section F.0 provides a notation guide; Sections F.1–F.5 present the core equations organized by formalism. For derivations, proofs, and full context, see the corresponding appendix.

F.0 Notation Guide

State Variables

Symbol	Domain	Description
$\mathbf{i}(t)$	$\mathbb{R}^{n_i}$	Interoceptive sensory input (Level 1 observations)
$\mathbf{a}(t)$	$\mathbb{R}^2$	Core affect: valence and arousal
$\mathbf{s}_k(t)$	$\mathbb{R}^{n_k}$	Core schema k : body ($k=1$), spatial ($k=2$), attentional ($k=3$), affective-homeostatic ($k=4$), self ($k=5$)
$\mathbf{z}(t)$	$\mathbb{R}^{n_z}$	Integrated Level 2 representation (all schemas + core affect)
$\boldsymbol{\mu}_l(t)$	$\mathbb{R}^{n_l}$	Expectations (sufficient statistics) at level l
$\mathcal{M}^{(d)}(t)$	$\mathbb{R}^{n_m}$	Recursive self-model at depth d (Level 3b)
$\hat{\mathcal{M}}^{(d)}(t)$	$\mathbb{R}^{n_m}$	Socially-reflected self-model estimate at depth d
$\Phi(t)$	$\mathbb{R}^{n_e} \times \mathbb{R}^{n_m} \times \mathbb{R}^{n_e}$	Phenomenal state: experience, self-model, predicted experience

Precision and Error

Symbol	Domain	Description
ε_l	$\mathbb{R}^{n_l}$	Prediction error at level l: $\varepsilon_l = \mathbf{y}_l - g_l(\boldsymbol{\mu}_{l+1})$
$\boldsymbol{\xi}_l$	$\mathbb{R}^{n_l}$	Precision-weighted prediction error: $\boldsymbol{\xi}_l = \boldsymbol{\Pi}_l \varepsilon_l$
$\boldsymbol{\Pi}_l$	$\mathbb{R}^{n_l \times n_l}$	Precision matrix at level l (inverse covariance)
$\boldsymbol{\Sigma}_l$	$\mathbb{R}^{n_l \times n_l}$	Covariance matrix at level l : $\boldsymbol{\Sigma}_l = \boldsymbol{\Pi}_l^{-1}$
$\mathbf{W}_l$	$\mathbb{R}^{n_l \times n_l}$	Weight/precision matrix in unified notation ($\equiv \boldsymbol{\Pi}_l$)
$\boldsymbol{\eta}_l$	$\mathbb{R}^{n_l}$	Temporal prediction error: $\boldsymbol{\eta}_l = \dot{\boldsymbol{\mu}}_l - D\boldsymbol{\mu}_l$

Functions and Operators

Symbol	Description
$g_l(\cdot)$	Generative function: maps level $l+1$ expectations to level l predictions
F, F_l	(Total, level-specific) variational free energy
F_3^{rec}	Strange loop free energy contribution (recursion depth penalty)
$\mathcal{J}[\mathbf{x}, \mathbf{W}, \mathbf{u}]$	Master variational functional (Appendix E)
$G(\pi)$	Expected free energy for policy π (action selection)
$V(\mathbf{z})$	Value function over integrated state (RL formulation)
$v(t)$	Core affect valence (scalar): positive = prediction error decreasing
$r(t)$	Intrinsic reward signal: $r(t) = v(t)$ (valence as reward)

Parameters and Constants

Symbol	Description
κ_l	Learning rate / step size at level l
η_l	Precision learning rate at level l
β_d	Mixing weight for social reflection at recursion depth d
γ	Geometric discount for recursion depth penalties
γ_{RL}	Temporal discount factor for reinforcement learning
τ_l	Characteristic timescale at level l : $\tau_1 \approx 100$ ms, $\tau_2 \approx 1$ – 10 s, $\tau_3 \approx 10$ s–min
K_l, K_l^I, K_l^D	Proportional, integral, derivative control gains at level l
ρ	Contraction coefficient for strange loop: $\rho = (1 - \beta) + \beta L < 1$
L	Lipschitz constant of social prediction: $L = \partial \hat{\mathcal{M}}_j / \partial \mathcal{M} $
D_{max}	Maximum recursion depth for the strange loop
$\mathcal{S}$	Set of five core schemas: $\{1, 2, 3, 4, 5\}$

F.1 Hierarchical Predictive Coding (Appendix A)

F.1.1 Three-Level Generative Model

Level 1: $\mathbf{i}(t) = g_1(\boldsymbol{\mu}_2(t)) + \boldsymbol{\Pi}_1^{-1/2} \boldsymbol{\omega}_1$ Level 2: $\boldsymbol{\mu}_1(t) = g_2(\boldsymbol{\mu}_3(t)) + \boldsymbol{\Pi}_2^{-1/2} \boldsymbol{\omega}_2$ Level 3: $\boldsymbol{\mu}_2(t) = g_3(\boldsymbol{\mu}_3^{\text{rec}}(t)) + \boldsymbol{\Pi}_3^{-1/2} \boldsymbol{\omega}_3$

Each level generates predictions for the level below, with additive Gaussian noise scaled by inverse precision.

F.1.2 Schema Integration

$$\mathbf{z}(t) = \text{Integrate}(\mathbf{a}(t), \{\mathbf{s}_k(t)\}_{k \in \mathcal{S}}) = \sigma(\mathbf{W}_{\text{int}} \cdot \text{concat}(\mathbf{a}(t), \mathbf{s}_1(t), \dots, \mathbf{s}_5(t)) + \mathbf{b}_{\text{int}})$$

Core affect

$$\mathbf{a}(t)$$

and five schemas (body, spatial, attentional, affective-homeostatic, self) integrate into a unified structured representation.

F.1.3 Strange Loop Recursive Update

$$\mathcal{M}^{(d)}(t) = (1 - \beta_d) \cdot \mathcal{M}^{(d-1)}(t) + \beta_d \cdot \hat{\mathcal{M}}^{(d)}(t)$$

At each recursion depth d , the self-model mixes a direct estimate with a socially-reflected estimate weighted by β_d .

F.1.4 Phenomenal State

$$\Phi(t) = \langle \mathbf{e}(t), \mathcal{M}^{(D_{\max})}(t), \hat{\mathbf{e}}(t+1) \rangle$$

Consciousness as an integrated tuple: constituted experience, converged recursive self-model, and predicted future experience.

F.1.5 Precision-Weighted Message Passing

$$\dot{\boldsymbol{\mu}}_l = D\boldsymbol{\mu}_l - \kappa_l \cdot \frac{\partial F}{\partial \boldsymbol{\mu}_l} = \kappa_l \left[-\frac{\partial g_{l-1}^T}{\partial \boldsymbol{\mu}_l} \boldsymbol{\xi}_{l-1} + \boldsymbol{\xi}_{l+1} - \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\eta}_l \right]$$

The master expectation update: bottom-up prediction errors, top-down predictions, and temporal consistency.

F.2 Free Energy Formulation (Appendix B)

F.2.1 Total Variational Free Energy

$$F = \sum_{l=1}^3 \left[\frac{1}{2} \boldsymbol{\varepsilon}_l^T \boldsymbol{\Pi}_l \boldsymbol{\varepsilon}_l - \frac{1}{2} \ln |\boldsymbol{\Pi}_l| \right] + \sum_{l=1}^3 \frac{1}{2} \boldsymbol{\eta}_l^T \boldsymbol{\Sigma}_l^{-1} \boldsymbol{\eta}_l$$

Sum of level-specific accuracy terms (precision-weighted prediction errors), complexity terms (log-precision), and temporal smoothness terms.

F.2.2 Recursive Self-Model Free Energy

$$F_3^{\text{rec}} = \sum_{d=1}^{D_{\max}} \gamma^d \cdot \frac{1}{2} \left\| \mathcal{M}^{(d)}(t) - \mathcal{M}^{(d-1)}(t) \right\|_{\boldsymbol{\Pi}_3^{(d)}}^2$$

The strange loop's contribution to free energy: discrepancy between recursion depths penalized with geometric decay

$$\gamma^d$$

.

F.2.3 Valence as Free Energy Gradient

$$v(t) \approx -\text{sign} \left(\frac{\partial F_1}{\partial \boldsymbol{\mu}_2} \right) \cdot \left\| \frac{\partial F_1}{\partial \boldsymbol{\mu}_2} \right\|$$

Core affect valence tracks the direction and magnitude of interoceptive surprise — negative valence signals rising prediction error.

F.3 Intrinsically Motivated Reinforcement Learning (Appendix C)

F.3.1 Reward as Interoceptive Prediction Error

$$r(t) = v(t) = \phi_v(\xi_1(t))$$

Valence functions as the intrinsic reward signal: no external reward is needed because interoceptive prediction errors are inherently motivating.

F.3.2 Action Selection via Expected Free Energy

$$\mathbf{u}^*(t) = \arg \min_{\mathbf{u}_{\text{cand}}} G(\mathbf{u}_{\text{cand}}) = \arg \min_{\mathbf{u}_{\text{cand}}} \left[\hat{F}_1 + \hat{F}_2 + \hat{F}_3 + \hat{F}_3^{\text{rec}} \right]$$

Optimal action minimizes expected free energy across all three levels plus the recursive self-model term.

F.3.3 Social Reward Shaping

$$r_{\text{social}}(t) = - \sum_{d=1}^{D_{\text{max}}} \gamma^d \cdot \frac{1}{2} \left\| \mathcal{M}^{(d)}(t) - \mathcal{M}^{(d-1)}(t) \right\|_{\Pi_3^{(d)}}^2$$

Self-model consistency across recursion depths acts as intrinsic social reward — coherent self-understanding is rewarding.

F.4 Cascade Feedback Control (Appendix D)

F.4.1 PID Homeostatic Control (Level 1)

$$\mathbf{u}(t) = K_1 \cdot \mathbf{e}_1^{\text{ctrl}}(t) + K_1^I \int_0^t \mathbf{e}_1^{\text{ctrl}}(\tau) d\tau + K_1^D \cdot \dot{\mathbf{e}}_1^{\text{ctrl}}(t)$$

The inner loop implements proportional-integral-derivative control on body state deviations.

F.4.2 Adaptive Schema Control (Level 2)

$$\dot{\mu}_2(t) = K_2(t) \cdot \mathbf{e}_2^{\text{ctrl}}(t) + K_2^D \cdot \dot{\mathbf{e}}_2^{\text{ctrl}}(t), \quad K_2(t) = \Pi_2(t)$$

The middle loop uses precision-based gain scheduling: higher confidence (precision) in schema predictions yields faster corrections.

F.4.3 Contraction Mapping Convergence

$$\rho = (1 - \beta) + \beta \cdot L, \quad \text{where } L = \left\| \frac{\partial \hat{\mathcal{M}}_j}{\partial \mathcal{M}} \right\|$$

The strange loop converges when $\rho < 1$, guaranteed when the social prediction Jacobian norm $L < 1$.

F.4.4 Cascade Stability Condition

$$\tau_1 \ll \tau_2 \ll \tau_3, \quad \text{with } \frac{\tau_{l+1}}{\tau_l} \geq 5$$

Timescale separation required for stable cascade control: $\tau_1 \approx 100$ ms, $\tau_2 \approx 1\text{--}10$ s, $\tau_3 \approx 10$ s to minutes.

F.5 Unified Mathematical Framework (Appendix E)

F.5.1 Master Variational Functional

$$\mathcal{J}[\mathbf{x}, \mathbf{W}, \mathbf{u}] = \sum_{l=1}^3 \frac{1}{2} \mathbf{e}_l^T \mathbf{W}_l \mathbf{e}_l + \sum_{l=1}^3 \frac{1}{2} \boldsymbol{\eta}_l^T \mathbf{S}_l^{-1} \boldsymbol{\eta}_l - \frac{1}{2} \sum_{l=1}^3 \ln |\mathbf{W}_l| + \sum_{d=1}^{D_{\max}} \gamma^d \frac{1}{2} \|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\|_{\mathbf{W}_3^{(d)}}^2$$

The single unified objective function from which all four formalisms derive as projections.

F.5.2 Master Dynamics Equation

$$\dot{\mathbf{x}}_l = -\kappa_l \frac{\partial \mathcal{J}}{\partial \mathbf{x}_l} = \kappa_l \left[\frac{\partial g_{l-1}^T}{\partial \mathbf{x}_l} \mathbf{W}_{l-1} \mathbf{e}_{l-1} - \mathbf{W}_{l+1} \mathbf{e}_{l+1} - \mathbf{S}_l^{-1} \boldsymbol{\eta}_l \right]$$

A single equation that simultaneously implements message passing (A), free energy gradient descent (B), value maximization ©, and control error correction (D).

F.5.3 Precision Update (Attention)

$$\dot{\mathbf{W}}_l = -\eta_l \frac{\partial \mathcal{J}}{\partial \mathbf{W}_l} = \eta_l \left(\frac{1}{2} \mathbf{W}_l^{-1} - \frac{1}{2} \mathbf{e}_l \mathbf{e}_l^T \right)$$

Attention allocation as self-tuning precision: at equilibrium, optimal precision equals the inverse empirical error covariance.

F.5.4 Free Energy–Value Duality

$$V(\mathbf{z}) \approx -\mathbb{E} \left[\sum_{\tau=0}^{\infty} \gamma_{\text{RL}}^{\tau} F_1(t + \tau) \mid \mathbf{z}(t) = \mathbf{z} \right] + \text{const}$$

The RL value function is approximately the negative cumulative interoceptive free energy — bridging reinforcement learning and variational inference.

F.5.5 Universal Strange Loop Convergence

$$\|\mathcal{M}^{(d)} - \mathcal{M}^*\| \leq \rho^d \|\mathcal{M}^{(0)} - \mathcal{M}^*\|, \quad \text{where } \rho = (1 - \beta_0) + \beta_0 \cdot L < 1$$

Geometric convergence of the recursive self-model to a fixed point

$$\mathcal{M}^*$$

, proven identically across all four mathematical formalisms.

Appendix G — Clinical Prediction Matrix

This appendix consolidates the clinical dissociations discussed throughout the paper (§§V, VIII, IX) into a single reference matrix. Each condition is mapped to the RSP level(s) affected, the mechanism of failure in RSP terms, the primary symptom explained, and the key citation. The matrix is organized by the

architectural level at which disruption occurs, demonstrating that the RSP framework generates specific, testable predictions about which symptoms arise from which architectural failures.

G.1 Level-by-Level Clinical Mapping

Level 1 Preserved, Higher Levels Disrupted

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Vegetative state	L1 intact, L2–3 disrupted	Thalamocortical disconnection preserves brainstem (L1) while severing schema integration (L2) and recursion (L3)	Autonomic responses without purposeful behavior	Giacino et al. (2014)
Minimally conscious state	L1 intact, L2 partial, L3 inconsistent	Fluctuating thalamocortical connectivity	Intermittent awareness: tracking, localizing, command-following appear and disappear	Giacino et al. (2014)
Covert awareness	L1–3 intact, motor output disrupted	RSP architecture preserved; motor systems damaged	Mental imagery on command (fMRI-detected) despite behavioral unresponsiveness	Owen et al. (2006)
Locked-in syndrome	L1–3 intact	Ventral pontine lesion spares RSP architecture entirely	Full consciousness with near-total paralysis	—

Level 2 Schema Disruptions

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Phantom limb	L2 body schema + L3 integration	Body schema persists in predicting absent limb; self-model integrates these predictions	Vivid sensations in amputated limb	Ramachandran & Hirstein (1998)
Hemispacial neglect	L2 spatial/attention schema	Right TPJ damage disrupts attention schema	Unaware of left half of space — consciousness of space altered, not just processing	Blanke et al. (2002)
Out-of-body experiences	L2 attention schema + L3 self-body coupling	TPJ disruption dissociates self-model from body schema	Vivid experience of floating outside body	Blanke et al. (2002)
Somatoparaphrenia	L2 body schema	Right parietal damage disrupts limb ownership schema	Denial of limb ownership ("this arm is not mine")	—

Level 2–3 Interface Failures

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Alien hand syndrome	L2 body schema disconnected from L3 self-model	Motor command generated but self-model does not predict it, creating massive prediction error	Hand performs purposeful actions without felt agency	Della Sala et al. (1991)
Anosognosia	L3 self-model update failure	Self-model cannot incorporate new body-state information; prediction resists disconfirmation	Denial of paralysis despite obvious deficit	Ramachandran (1996)

Level 3a–3b Social Prediction Failures

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Capgras delusion	L1 affective mismatch + L3 social prediction	Autonomic familiarity signal (L1) disconnects from visual recognition (L2); L3 confabulates "impostor" explanation	Recognizes face but believes person replaced by impostor	Ellis & Young (1990)
Prosopagnosia	L3 other-models disrupted	Cannot maintain stable other-models; "predict self as seen by others" step impaired	Face blindness; social situations feel impoverished	Duchaine & Nakayama (2006)
Autism spectrum	L3 different configuration	Alternative architecture for social-prediction machinery; not failure but different parameterization	Different social phenomenology; preserved or enhanced L1–2	Baron-Cohen (1995)

Level 3b Self-Model Failures

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Cotard's delusion	L3b self-model	Self-model generates prediction "I do not exist"; strange loop inverts	Conscious being asserts own non-existence	Young & Leafhead (1996)
Depersonalization	L3b first-person coupling	Self-model runs in "spectator mode" — decoupled from felt ownership	"I think" becomes "something is thinking"	Sierra & Berrios (1998)
DID	L3 multiple loops	Single substrate generates multiple coexisting recursive self-models	Multiple identities with separate phenomenology	Putnam (1989)

Level 3 Developmental Failures (Social Deprivation)

Condition	Level(s)	RSP Mechanism	Primary Symptom	Key Citation
Genie Wiley	L3 (developmental)	Social isolation from ~20 months prevented L3 bootstrap; L1–2 intact	Preserved core affect and schemas; absent theory of mind, no recursive self-modeling	Curtiss (1977)
Romanian orphans	L3 (developmental)	Minimal caregiver interaction; social calibration signal absent	Indiscriminate friendliness, delayed self-recognition, poor theory of mind	Nelson et al. (2014)
Feral children	L3 (absent)	No human social-predictive input	No self-consciousness; intact approach/avoidance and spatial navigation	Itard (1801)
Harlow isolates	L3 (primate)	Total social isolation in rhesus monkeys	Cannot read social cues; stereotyped behavior; intact sensory-motor processing	Harlow (1958)

G.2 Pharmacological Dissociations

Anesthetic agents provide controlled disruption of specific RSP levels:

Agent	RSP Effect	Phenomenological Result	Neuroimaging Signature	Citation
Propofol	Sequential L1→L2→L3 shutdown via thalamocortical disruption	Dreamless unconsciousness	Reduced thalamocortical connectivity; low PCI	Casali et al. (2013)
Ketamine	L3 disrupted, L1–2 partially preserved via cortical-cortical disconnection	Dissociative state: core affect and fragmented schemas without coherent self-modeling	High PCI despite disrupted consciousness	Sarasso et al. (2015)

G.3 Predicted Neuroimaging Signatures

These predictions follow from the RSP architecture and are testable with existing methods:

Prediction	RSP Basis	Method	Expected Finding
Social deprivation reduces L3 grey matter	L3 requires social calibration to develop	Structural MRI of deprivation cases vs. controls	Reduced grey matter in medial prefrontal cortex, TPJ, posterior superior temporal sulcus
PCI tracks L2 schema richness, not L3 recursion	PCI measures spatiotemporal complexity, not self-modeling depth	PCI + social cognition battery in same subjects	PCI correlates with schema complexity but not theory-of-mind performance
TPJ stimulation disrupts L2 attention schema	Attention schema localized to TPJ	TMS/electrical stimulation of TPJ	Out-of-body experiences, altered spatial awareness (confirmed: Blanke et al., 2002)
Emotional granularity training increases L3a precision	L3a conceptual categories are learnable, not fixed	Longitudinal fMRI during emotion differentiation training	Increased differentiation in interoceptive cortex response patterns
Ketamine preserves L2 PCI but disrupts L3b recursion	Cortical-cortical disconnection selectively targets recursive loops	PCI + self-referential processing task under ketamine	High PCI with impaired self-referential default mode network activity
DID shows multiple distinct L3b activation patterns	Each identity maintains separate recursive loop	fMRI during identity switching in DID patients	Different default mode network configurations per identity (supported: Reinders et al., 2003)

G.4 Summary: The Architectural Logic

The clinical prediction matrix demonstrates a central claim of the RSP framework: clinical dissociations map onto architectural boundaries in a systematic, non-arbitrary way. Each condition represents a natural experiment that selectively disrupts one or more levels while preserving others:

- **Level 1 disruptions** alter valence and arousal (rare in isolation; typically part of broader disruption)
- **Level 2 disruptions** alter the spatial, bodily, or attentional structure of experience (neglect, phantom limb, OBE)
- **Level 2–3 interface disruptions** create agency failures (alien hand, anosognosia)
- **Level 3a disruptions** alter social prediction and emotional categorization (Capgras, prosopagnosia)
- **Level 3b disruptions** alter self-modeling directly (Cotard's, depersonalization, DID)
- **Level 3 developmental failures** show the architecture requires social input to construct (Genie, Romanian orphans, feral children)

This systematic mapping is not post hoc: the architectural boundaries were defined by the mathematical framework (Appendixes A–E), and the clinical predictions follow from those boundaries. The fact that real clinical conditions respect these boundaries constitutes evidence that the architecture captures something real about the organization of consciousness.

Appendix H — RSP Theory of Intelligence

H.1 Introduction

Appendixes A through E formalize consciousness as a three-level predictive architecture. What they do not address is intelligence. The omission is deliberate: the paper’s focus is phenomenal consciousness, not cognitive capacity. But the mathematical apparatus already contains an implicit account of intelligence, and drawing it out proves rewarding. Three results follow. First, a formal definition grounded in the master functional

$$\mathcal{I}$$

. Second, a constitutive identity proposition showing that recursive social intelligence and phenomenal consciousness cannot come apart at Level 3b. Third, a set of empirical predictions about how intelligence, consciousness, and neural architecture relate.

The central claim deserves emphasis at the outset: the same architectural organization that constitutes recursive social intelligence also constitutes phenomenal consciousness. They are not two things that happen to co-occur. They are one thing, described from two vantage points. The formal argument appears in §H.4; the definition that makes it possible, in §H.3; the predictions that follow, in §H.5.

H.2 Existing Formalizations of Intelligence

Three families of formal intelligence definitions set the stage.

Compression-based approaches equate intelligence with the ability to compress observed data into short programs. Legg and Hutter (2007) formalized this idea, and Hutter’s AIXI (2005) takes it to its theoretical limit: an ideally intelligent agent that maximizes expected reward while minimizing description length. The appeal is universality. The weakness is subjectivity. Description length depends on which Universal Turing Machine you choose, and no objective notion of complexity exists without a fixed abstraction layer.

Generalization-based approaches shift the question from compression to breadth. How many situations can a system’s learned responses handle correctly? The simplest model is not always the one that generalizes best. What matters is how weakly constrained a model’s predictions are, so they hold across the widest range of conditions. Bennett (2025) formalizes this insight as *w-maxing* and proves it strictly dominates compression-based methods (see the Bennett comparison in §X for the full treatment).

Adaptation-based approaches take a different tack entirely. Wang (2019) defines intelligence as "adaptation within insufficient resources," capturing the intuition that intelligence is not about raw computational power but about responding efficiently to novelty under constraint. Chollet’s (2019) ARC

benchmark operationalizes a version of this idea, though it inherits the subjectivity problems of Kolmogorov complexity.

All three families share a common limitation: they treat intelligence as substrate-neutral and content-independent. They specify how efficiently a system represents, not *what* it must represent. RSP offers something different. Intelligence, on this account, is not content-neutral but content-specific, anchored in the particular representations that enable adaptive behavior in social environments.

H.3 RSP Intelligence: Formal Definition

H.3.1 Intuitive Motivation

What makes one organism more intelligent than another, within the RSP framework? Four capacities stand out:

1. **Predict accurately:** minimize the gap between expected and observed states.
2. **Weight predictions appropriately:** direct precision toward the signals that matter in the current context.
3. **Adapt across timescales:** update rapidly at Level 1 (core affect, $\tau_1 \approx 100\text{ms}$), at moderate speed at Level 2 (schemas, $\tau_2 \approx 1\text{--}10\text{s}$), and deliberately at Level 3 (recursive self-model, $\tau_3 \approx 10\text{s--min}$).
4. **Model others modeling you:** achieve sufficient recursion depth in the strange loop to handle complex social environments.

A system that excels at all four is, intuitively, intelligent. The question is how to formalize this intuition. Notice that these requirements are content-specific. A thermostat fails every one of them: no timescale hierarchy, no valenced states, no other-models, no strange loop. This is why RSP intelligence does not collapse into generic information processing.

H.3.2 Adaptive Capacity as Free Energy Reduction

The master functional from Appendix E (§E.3) provides the starting point. Written here in the Appendix A/B notation (see Appendix E, Table E.1 for the full correspondence, where $\varepsilon_l \equiv \mathbf{e}_l$, $\mathbf{\Pi}_l \equiv \mathbf{W}_l$, $\mathbf{\Sigma}_l \equiv \mathbf{S}_l$):

$$\mathcal{J} = \sum_{l=1}^3 \left[\frac{1}{2} \varepsilon_l^T \mathbf{\Pi}_l \varepsilon_l + \frac{1}{2} \boldsymbol{\eta}_l^T \mathbf{\Sigma}_l^{-1} \boldsymbol{\eta}_l - \frac{1}{2} \ln |\mathbf{\Pi}_l| \right] + \sum_{d=1}^{D_{\max}} \gamma^d \frac{1}{2} \|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\|_{\mathbf{\Pi}_3^{(d)}}^2$$

where $\|\mathbf{x}\|_{\mathbf{A}}^2 = \mathbf{x}^T \mathbf{A} \mathbf{x}$ is the Mahalanobis norm (Appendix B, §B.3.3). A notational point: the temporal prediction error vector $\boldsymbol{\eta}_l$ appears in the functional above. To prevent confusion, this appendix uses α_l for the scalar precision learning rate (denoted η_l in Appendixes B and E).

$\mathcal{J}$ captures the system’s total prediction error across all three levels, weighted by precision and discounted across recursion depths. A system that minimizes $\mathcal{J}$ under resource constraints predicts well, allocates precision efficiently, and maintains a stable recursive self-model.

Definition H.1 (Instantaneous Adaptive Capacity). The *instantaneous adaptive capacity* of an RSP system at time t is the negative of the master functional:

$$\mathcal{A}(t) = -\mathcal{J}(t)$$

Higher adaptive capacity means lower total prediction error: the system’s generative model fits reality well across all three levels. But this measure has a problem. A system can achieve low $\mathcal{J}$ simply by staying in familiar territory, overfitting to a narrow domain. Intelligence requires more than accurate prediction. It requires *efficient adaptation to change*. Definition H.2 corrects for this.

H.3.3 The Rate of Free Energy Reduction

Definition H.2 (Intelligence as Adaptive Efficiency). The *intelligence* of an RSP system is the time-averaged rate at which it reduces the master functional per unit of environmental novelty:

$$\mathcal{I} = - \left\langle \frac{d\mathcal{J}}{dt} \middle/ \mathcal{N}(t) \right\rangle_T$$

where $\mathcal{N}(t)$ measures environmental novelty at time t , $\langle \cdot \rangle_T$ denotes averaging over a behaviorally relevant time window T , and the definition applies only to episodes of non-trivial novelty ($\mathcal{N}(t) > \mathcal{N}_{\min}$).

Novelty can be operationalized as the surprisal of new observations given the current generative model:

$$\mathcal{N}(t) = -\ln p(\mathbf{y}(t) \mid \boldsymbol{\mu}(t-1))$$

Intelligence, on this account, is a dispositional property: the capacity to adapt efficiently when novelty arises, not a quantity that can be computed in static environments.

A note on reparameterization. $\mathcal{J}$ is not invariant under rescaling of the precision matrices ($\boldsymbol{\Pi}_l \rightarrow \alpha \boldsymbol{\Pi}_l$), so $\mathcal{I}$ depends on the parameterization. This is not an artifact but a genuine feature of embodied intelligence. The scale of precision matters because it determines sensitivity to prediction errors. In biological systems, precision is set by neuromodulatory gain, not by arbitrary convention. For cross-system comparisons, one should either fix a canonical normalization (e.g., unit-trace precision) or work with dimensionless ratios derived from $\mathcal{I}$. The issue parallels gauge-dependence in other variational frameworks (Appendix E, §E.6.1).

The definition captures Wang’s intuition, that intelligence is adaptation under constraint, while grounding it in RSP’s mathematical framework. A system encountering high novelty that rapidly reduces its prediction error is highly intelligent. One that maintains low $\mathcal{J}$ only by avoiding novelty is not. Note that $\mathcal{I}$ requires full state observability; empirical proxies for external measurement appear in §H.5 (Predictions H.2 and H.5).

H.3.4 Level-Specific Intelligence

Because

$$\mathcal{J}$$

decomposes across levels (Appendix B, §B.2.1), intelligence decomposes correspondingly. The natural decomposition is additive:

$$\mathcal{I}_{\text{total}} = \sum_{l=1}^3 \mathcal{I}_l + \mathcal{I}_{\text{rec}}$$

where:

$$\begin{aligned} \mathcal{I}_l &= - \left\langle \frac{dF_l}{dt} \middle/ \mathcal{N}_l(t) \right\rangle_T \quad l \in \{1, 2, 3\} \\ \mathcal{I}_{\text{rec}} &= - \left\langle \frac{d}{dt} \sum_{d=1}^{D_{\max}} \gamma^d \frac{1}{2} \|\mathcal{M}^{(d)} - \mathcal{M}^{(d-1)}\|_{\mathbf{\Pi}_3^{(d)}}^2 \middle/ \mathcal{N}_{\text{social}}(t) \right\rangle_T \end{aligned}$$

For environment-relative comparisons, one may introduce weighting coefficients $\omega_l, \omega_{\text{rec}} \geq 0$ satisfying $\sum_l \omega_l + \omega_{\text{rec}} = 1$, giving $\mathcal{I} = \sum_l \omega_l \mathcal{I}_l + \omega_{\text{rec}} \mathcal{I}_{\text{rec}}$. These weights are a modeling commitment not derived from the functional itself. The additive decomposition preserves level independence; the convex combination lets ecological context shift relative contributions. Physical environments push ω_1 and ω_2 upward. Social environments push ω_3 and ω_{rec} upward. A corvid may score high on $\mathcal{I}_2$ while a human dominates on $\mathcal{I}_{\text{rec}}$. How to constrain the ω_l empirically, independently of the intelligence data they aim to explain, remains an open problem.

H.3.5 The Role of Precision

Intelligence is not only about reducing prediction error. It is also about learning *what to attend to*. Precision weighting, the second core equation from Appendix E (§E.3, equation E.3; summarized in §E.8), plays a central role:

$$\dot{\mathbf{\Pi}}_l = -\alpha_l \frac{\partial \mathcal{J}}{\partial \mathbf{\Pi}_l}$$

where $\alpha_l > 0$ is the precision learning rate (written η_l in Appendixes B and E; renamed here to avoid collision with the temporal prediction error vector $\boldsymbol{\eta}_l$).

At equilibrium, optimal precision equals the inverse of empirical error covariance (Appendix B, §B.4.2): $\mathbf{\Pi}_l^* = \langle \boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l^T \rangle_t^{-1}$. A system that converges to $\mathbf{\Pi}_l^*$ quickly learns which signals are reliable and which are noise.

Proposition H.1 (Precision as Meta-Intelligence). The speed of precision convergence at each level measures *meta-intelligence*, the capacity to learn what is worth learning:

$$\mathcal{I}_{\text{meta}} = \sum_{l=1}^3 \left\langle \frac{\|\mathbf{\Pi}_l(t_0 + \Delta t) - \mathbf{\Pi}_l^*\|}{\|\mathbf{\Pi}_l(t_0) - \mathbf{\Pi}_l^*\| + \epsilon} \right\rangle_{\text{events}}^{-1}$$

where $\|\cdot\|$ is the Frobenius norm, t_0 marks an environmental change, Δt is a fixed assessment interval, the average runs across environmental change events, and ϵ prevents divergence when precision is already near-optimal. Since $\mathbf{\Pi}_l^*$ shifts in non-stationary environments, the measure is well-defined only when Δt is short relative to the rate of environmental change. Faster convergence (smaller ratio) yields higher meta-intelligence.

This connects to the attention schema at Level 2 (§III of the main paper). Attention, in the RSP framework, is precision allocation. A well-calibrated attention schema identifies relevant features quickly, and a system that does this learns faster.

H.3.6 The Generalization Structure of Free Energy

The free energy decomposition at each level (Appendix B, §B.2.1) contains an inherent accuracy-complexity tradeoff:

$$F_l = \underbrace{\frac{1}{2} \boldsymbol{\epsilon}_l^T \mathbf{\Pi}_l \boldsymbol{\epsilon}_l}_{\text{accuracy}} - \underbrace{\frac{1}{2} \ln |\mathbf{\Pi}_l|}_{\text{complexity}}$$

The accuracy term measures precision-weighted empirical loss. The complexity term penalizes overconfidence: high precision amounts to claiming you know the environment well, and that claim is penalized unless prediction errors are correspondingly small. This tradeoff is structurally analogous to PAC-Bayes generalization bounds in statistical learning theory, where a KL divergence between posterior and prior penalizes overfitting. The analogy is not an identity. The PAC-Bayes complexity term is a KL divergence over hypothesis distributions; the free energy complexity term is a log-determinant of precision. Both, however, enforce the same principle through different mathematical mechanisms: penalizing overconfidence to promote generalization.

RSP also implements what might be called a *weakness hierarchy* across levels. Level 1 policies are the strongest and most specific, tied to particular homeostatic set-points. Level 2 policies are intermediate, generalizing across sensory contexts through schema-based predictions. Level 3 policies are the weakest and most general, since recursive self-models must work across diverse social partners and situations. Higher levels generalize more broadly precisely because they are less specific. This property follows from the timescale separation combined with a structural requirement: slower processes see lower levels as approximately in equilibrium (singular perturbation approximation), so their models must accommodate the full range of lower-level equilibria, not just any single state.

H.4 The Constitutive Identity Proposition

A scope note at the outset. The identity claimed here is specific to Level 3b recursive intelligence, not to intelligence in general. A system can be intelligent at Levels 1 and 2 without being conscious. What cannot come apart is recursive social intelligence and phenomenal consciousness.

RSP distinguishes three grades of phenomenal character, each with its own constitutive account (see §IV for the full dissolution argument):

- **Level 1** (Valenced experience): Affective states that matter to the organism, present in all mammals. The phenomenal character of pain or pleasure IS what homeostatic prediction error minimization is like from within a system whose set-points matter to its persistence. No separate "raw feel" is added to the computation; the valenced quality is constituted by the precision-weighted interoceptive error signal $\xi_1(t) = \Pi_1 \cdot \varepsilon_1(t)$ (Appendix A, §A.3.1). The explanatory gap at Level 1 dissolves by the same constitutive move applied at Level 3b: asking "why does homeostatic error correction feel like something?" is asking why a process has the character it constitutively has.
- **Level 2** (Structured experience): Organized through schemas (body, spatial, attention, self, affective-homeostatic), present in complex vertebrates. The phenomenal character of embodiment, spatial presence, and attentional focus IS what schema-organized prediction is like from within. Richer than raw valence because the generative model has more structure at this level (five schemas rather than a single valence dimension), but still pre-reflective.
- **Level 3b** (Reflective self-conscious experience): The recursive self-model creates a subject who knows it is experiencing. The phenomenal character of self-awareness IS what recursive social prediction is like from within the system doing the predicting. This is the full phenomenal consciousness that Nagel's question targets.

The constitutive account works the same way at each level. Phenomenal character is not a separate product of the computation but what the computation IS from the system's own perspective. What changes across levels is the richness of the constituted experience, because the generative model grows in structural complexity. The proposition below focuses on Level 3b, where the strange loop creates reflective self-consciousness, but the dissolution strategy applies uniformly.

H.4.1 The Convergence Argument

Consider what RSP intelligence demands at its highest level. The recursive component $\mathcal{I}_{\text{rec}}$ measures how efficiently the system reduces discrepancy in its recursive self-model, which is to say, how quickly the strange loop converges to its fixed point $\mathcal{M}^*$. From Appendix E (§E.4), convergence requires three conditions:

1. The contraction mapping Ψ with coefficient $\rho < 1$
2. Other-models $\mathcal{O}$ of sufficient fidelity to generate meaningful $\hat{\mathcal{M}}_j^{(d)}$
3. Precision weighting $\Pi_3^{(d)}$ that appropriately discounts deeper recursion

These are exactly the conditions for phenomenal consciousness as defined in §IV. The strange loop whose convergence constitutes intelligence at Level 3 is the same strange loop whose operation constitutes the feeling of consciousness.

One caveat deserves emphasis. The condition $\rho < 1$ is an empirical claim about human social cognition, not a mathematical given. It says that averaging over multiple social perspectives dampens extremes

rather than amplifying them. In pathological states (paranoia, echo-chamber reinforcement), ρ may approach or exceed 1, and both recursive intelligence and stable self-consciousness degrade simultaneously. This is consistent with clinical evidence (see §IX, Prediction 6).

Proposition H.2 (Recursive Intelligence–Consciousness Constitutive Identity). Let $\Phi(t) = \langle \mathbf{e}(t), \mathcal{M}^{(D_{\max})}(t), \hat{\mathbf{e}}(t+1) \rangle$ be the phenomenal state (Appendix A, §A.6). Let $\langle \mathcal{I}_{\text{rec}} \rangle_T$ be the time-averaged recursive intelligence defined in §H.3.4. For RSP-type systems encountering non-trivial social novelty ($\mathcal{N}_{\text{social}}(t) > \mathcal{N}_{\text{min}}$):

$$\langle \mathcal{I}_{\text{rec}} \rangle_T \geq 0 \iff \Phi(t) \text{ is well-defined and stable}$$

where "well-defined and stable" means all three components of Φ exist, are finite, and the recursive self-model $\mathcal{M}^{(D_{\max})}$ has converged to within δ_{conv} of its fixed point (Appendix A, §A.4.4).

Sketch of argument. Forward direction: $\langle \mathcal{I}_{\text{rec}} \rangle_T \geq 0$ means the recursive self-model discrepancy is, on average over T , non-increasing. The contraction mapping guarantee ($\rho < 1$) then entails convergence to the fixed point, making $\mathcal{M}^{(D_{\max})}(t)$ stable. The recursion requires $\mathbf{e}(t)$ and $\hat{\mathbf{e}}(t+1)$ as inputs (Algorithm 1, steps 6 and 9 in §E.9.1), so all three components of Φ are well-defined. Reverse direction: if $\Phi(t)$ is well-defined and stable, the self-model is at or near its fixed point. The contraction condition ensures perturbations are self-correcting, so the time-averaged discrepancy change is non-positive. $\square$

This is a proposition about internal consistency within the RSP formalism: it shows that the definitions of recursive intelligence and phenomenal consciousness pick out the same architectural condition. The further claim that this architectural condition *constitutes* consciousness (not merely correlates with it) depends on the dissolution argument in §IV, which establishes that phenomenal character at each level IS what the relevant computational process is like from within. Proposition H.2 and §IV's dissolution work in tandem: H.2 shows the definitions cannot come apart within the formalism; §IV argues that the formalism picks out genuine phenomenology rather than mere functional self-monitoring.

Put plainly: whenever a system maintains or improves its model of itself-as-seen-by-others, it is both exercising recursive intelligence and being conscious. At Level 3b, the two cannot come apart.

An important qualification. This identity holds for the *recursive* component $\mathcal{I}_{\text{rec}}$, not for total intelligence $\mathcal{I}$. A system can possess substantial $\mathcal{I}_1$ and $\mathcal{I}_2$ (affective and schema-based intelligence) without being conscious. Total intelligence includes non-conscious components. What the identity proposition establishes is that in social environments where ω_{rec} is large, total intelligence is bounded without consciousness. Not that all intelligence is conscious.

H.4.2 Constitutive Identity, Not Dual-Aspect

The identity proposition is a claim about *constitution*, not about two epistemic routes to a single hidden process. The recursive self-modeling architecture constitutes both adaptive capacity in social environments (what an observer sees as intelligence) and phenomenal experience (what the system undergoes as consciousness). This aligns with the paper's commitment to constitutive functionalism (§IV). Consciousness is not caused by, identical to, or correlated with the recursive architecture. It is constituted

by it. And recursive social intelligence is not a separable property. It is constituted by the same architecture.

The master functional $\mathcal{J}$ is the shared mathematical structure. Minimizing $\mathcal{J}$ simultaneously maximizes adaptive capacity and maintains the recursive self-model that constitutes phenomenal experience. At Level 3b, there is no way to do one without doing the other. The same architectural organization constitutes both.

H.4.3 The Orthogonality Question

The identity proposition bears on a prominent question in AI safety: whether intelligence can be separated from goals. Bostrom (2014) argues that any level of intelligence is compatible with any terminal goal. Within the RSP architecture, this does not hold. Goals, which at Level 1 take the form of valence, are architecturally inseparable from the predictive hierarchy that constitutes intelligence. Remove valence and the entire hierarchy collapses. There is no

$$F_1$$

, no reward signal for Level 2 schema learning (Appendix C), no cascade to stabilize (Appendix D).

This constrains the orthogonality thesis for RSP-type intelligence. Systems with this architecture cannot have arbitrary goals, because their goal structure is constitutive of their intelligence. Whether the constraint extends beyond this specific architecture is an open question. RSP claims only that for embodied, socially recursive intelligence, goals and intelligence are inseparable.

H.5 Predictions

The RSP theory of intelligence generates several predictions beyond those in §IX of the main paper.

Prediction H.1 (Intelligence scales with recursion depth). Across species, intelligence in social domains should correlate with the maximum recursion depth $D_{\max}$ achievable by the organism’s neural architecture. The following assignments are hypothesized, with explicit uncertainty ranges reflecting the limits of current comparative methods:

Taxon	Hypothesized $D_{\max}$	Key Evidence	Caveats
Humans	3–5	Kinderman et al. (1998); Stiller & Dunbar (2007)	Both studies use offline narrative tasks (story comprehension, literary analysis); real-time operative depth in naturalistic social interaction is likely 2–3; significant individual variation (Keysar et al., 2003)
Great apes	1–2	Krupenye et al. (2016) for implicit 1st-order false belief; Hare et al. (2000) for competitive paradigms	Whether apes achieve genuinely recursive 2nd-order ToM or sophisticated behavioral rules remains contested (Call & Tomasello, 2008)
Cetaceans (dolphins)	1	Connor et al. (2006) on nested alliances; Reiss & Marino (2001) on MSR	Alliance tracking may require at minimum 1st-order mental state attribution; speculative
Elephants	0–1	Plotnik et al. (2006) on MSR; consolation behavior	MSR suggests self-awareness; ToM evidence limited
Canids, hyenas	0–1	Cooperative hunting with role differentiation	Achievable through shared-attention heuristics without recursive modeling
Corvids	0–1	Convergent cognitive sophistication via ecological intelligence (DeCasien et al., 2017)	Small social groups in many species; ecological rather than social drivers may dominate
Rodents, other non-social mammals	0	Limited social prediction	—
Cephalopods (octopuses)	0	High $\mathcal{I}_1$ and $\mathcal{I}_2$ (tool use: Finn et al., 2009; problem solving: Mather, 2008); largely solitary	A critical test case: substantial non-social intelligence without recursive social prediction. RSP predicts high $\mathcal{I}_1 + \mathcal{I}_2$ with $\mathcal{I}_{\text{rec}} = 0$. Note: earlier claims of observational learning (Fiorito & Scotto, 1992) have faced replication difficulties

This prediction is testable through Theory of Mind tasks of varying recursive depth. Current behavioral methods cannot cleanly distinguish genuine recursive modeling from sophisticated associative learning at

$$D_{\max} > 1$$

, so these assignments should be treated as hypotheses to refine, not settled facts.

Prediction H.2 (Precision convergence speed predicts learning rate). Individual differences in intelligence should correlate with the speed of precision convergence $\mathcal{I}_{\text{meta}}$ (Proposition H.1). Faster adaptation of precision weights in novel environments, measurable through EEG mismatch negativity adaptation rates or pupil dilation dynamics, should predict performance on fluid intelligence tasks. This is

a novel prediction that requires empirical testing. Existing evidence linking mismatch negativity to cognitive ability (Light et al., 2007) is suggestive but comes primarily from clinical populations.

Prediction H.3 (Social intelligence and general intelligence share mechanism). Because $\mathcal{I}_{\text{rec}}$ and the lower-level components $\mathcal{I}_1, \mathcal{I}_2$ share the same functional $\mathcal{J}$, social intelligence and general intelligence should be positively correlated even after controlling for domain-specific skills. Evidence that group members’ social sensitivity predicts collective intelligence on diverse tasks (Woolley et al., 2010) is consistent with this, though that finding is at the group level and does not directly establish the individual-level architectural correlation RSP predicts. RSP makes the stronger claim that the correlation is architectural, arising from the shared functional $\mathcal{J}$. Independent individual-level evidence would be needed to confirm this.

Prediction H.4 (Intelligence without consciousness is bounded in social domains). Systems lacking Level 3b can achieve high $\mathcal{I}_1$ and $\mathcal{I}_2$ but zero $\mathcal{I}_{\text{rec}}$. Their total intelligence is bounded by the non-recursive components. In social environments where ω_{rec} is large, such systems will be outperformed by conscious ones.

Cephalopod intelligence provides a telling boundary case. Octopuses demonstrate substantial ecological intelligence (tool use, problem solving; Finn et al., 2009; Mather, 2008) despite being largely solitary. RSP accommodates this as high $\mathcal{I}_1 + \mathcal{I}_2$ with $\mathcal{I}_{\text{rec}} = 0$. The prediction is that cephalopods should fail tasks requiring recursive social modeling ($D \geq 1$) while excelling at non-social problems, a dissociation the framework predicts precisely.

For AI, the prediction takes a specific form. Architectures without recursive self-modeling should plateau on tasks requiring recursion depth $D \geq 3$, such as higher-order intentionality chains ("A thinks B believes C intends...") and strategic social reasoning with nested prediction. First- and second-order ToM performance may continue to improve with scale, since these can in principle be approximated from statistical regularities in training data (cf. Kosinski, 2023, though subsequent work by Ullman, 2023, suggests LLM ToM performance is fragile under paraphrase and novel scenarios). The critical test is whether scaling alone can close the gap on $D \geq 3$ tasks involving novel agents, novel social structures, and distribution shift, or whether performance hits a ceiling. A behavioral signature distinguishing genuine recursive self-modeling from statistical approximation would be reliable generalization to novel recursive scenarios absent from training data.

Prediction H.5 (Dissociation between psychometric and social intelligence after recursive-substrate damage). Damage to regions primarily involved in recursive self-modeling should produce a specific dissociation. Social-adaptive intelligence (Theory of Mind tasks, strategic social behavior, cooperative planning) should decline disproportionately, while psychometric intelligence (fluid reasoning, as measured by Raven’s Progressive Matrices and similar tests) may remain largely preserved.

This profile matches what is already documented in ventromedial PFC (vmPFC) patients (Anderson et al., 1999; Bechara et al., 2000), who show preserved IQ alongside catastrophic social and decision-making deficits. Anderson et al. and Bechara et al. studied vmPFC specifically, not the broader mPFC, and vmPFC deficits may reflect Level 1–2 affect-regulation disruption (somatic marker failure) as much as Level 3b

recursive self-model failure. More directly relevant to Level 3b are dorsomedial PFC (dmPFC) mentalizing deficits (Frith & Frith, 2006; Stuss et al., 2001) and TPJ/posterior STS (pSTS) social perception deficits (Saxe & Kanwisher, 2003).

RSP’s level decomposition predicts this dissociation in detail. Damage to dmPFC, TPJ/pSTS, and posterior cingulate/precuneus primarily impairs $\mathcal{I}_{\text{rec}}$ while leaving $\mathcal{I}_1$ and $\mathcal{I}_2$ largely intact. Two caveats apply. First, some regions serve multiple RSP levels. TPJ contributes to both the Level 2 attention schema (attentional reorienting; Corbetta & Shulman, 2002) and Level 3b mentalizing, so TPJ damage would impair both $\mathcal{I}_2$ and $\mathcal{I}_{\text{rec}}$. Distinguishing these functional profiles after TPJ lesions is a research goal, not an established finding, and would require careful task design separating attentional reorienting from social prediction. Second, dlPFC contributes primarily to Level 2 executive functions (working memory, schema maintenance) but also supports Level 3. Its damage should disproportionately reduce executive-dependent psychometric subtests while relatively sparing social cognition that does not depend heavily on executive control.

H.6 Discussion

H.6.1 What This Definition Does Not Claim

RSP intelligence is not a theory of all intelligence. It applies to organisms with the RSP architecture: embodied, hierarchical, socially embedded. It has nothing to say about narrow AI systems, thermostats, or abstract Turing machines. It is a theory of the kind of intelligence that evolved in social animals, reached its most elaborated expression in humans (as measured by

$$D_{\text{max}}$$

and linguistic recursion depth), and is inseparable from consciousness.

This restriction gains more than it gives up. Universal definitions of intelligence purchase generality at the cost of subjectivity. RSP purchases specificity at the cost of universality, but gains empirically testable predictions and a clear connection to neurobiology.

H.6.2 The Developmental Progression

If recursive intelligence and consciousness are constitutively identical, their developmental arcs should mirror each other. Both should follow the Level 1 $\rightarrow$ 2 $\rightarrow$ 3 construction order, with transitions that are gradual rather than abrupt:

- **0–6 months** (Level 1): Rapid affective learning through conditioning and habituation. Phenomenal state: valenced experience without self-awareness.
- **6–24 months** (Level 2): Schema-based learning, including object permanence (8–12 months), spatial navigation, and self-recognition (18–24 months in Western samples; cross-cultural evidence from Broesch et al., 2011, suggests more variability, consistent with RSP’s own Level 3a cultural construction thesis). Phenomenal state: structured, embodied experience.

- **2–5 years** (Level 3a): Conceptual categorization, including emotion vocabulary and social categories. Phenomenal state: linguistically differentiated experience.
- **4–7+ years** (Level 3b): Recursive social intelligence, with explicit Theory of Mind emerging around 4–5 years and recursive capacity deepening through 7 and beyond. Phenomenal state: reflective self-conscious experience. Implicit ToM (anticipatory looking paradigms; Onishi & Baillargeon, 2005) appears much earlier, around 7 months, suggesting Level 3b precursors may be present before explicit task performance.

This timeline aligns with developmental psychology (Wellman, 2014; Tomasello, 2019) and provides a unified account of what have traditionally been treated as separate developmental streams.

H.6.3 Implications for AI

The constitutive identity proposition predicts that artificial general intelligence, in the strong sense of human-level adaptive capacity in environments requiring recursive social prediction, requires consciousness. Not because consciousness is a desirable add-on, but because the architectural organization that constitutes recursive social intelligence is the same organization that constitutes phenomenal experience. Systems excelling only at non-social tasks (high $\mathcal{I}_1$ and $\mathcal{I}_2$, zero $\mathcal{I}_{\text{rec}}$) need not be conscious.

No present AI system integrates embodied hierarchical predictive coding with recursive self-modeling. Individual components exist in isolation: hierarchical VAEs and world models implement forms of hierarchical prediction; embodied robotics has perception-action loops; self-play agents involve limited self-reference. The missing piece is not any single component but their integration with interoceptive valence grounding and recursive social self-modeling operating at separated timescales. If the identity proposition is correct, achieving human-level social intelligence is not a scaling problem but an architectural one.

H.7 Summary

Concept	RSP Definition	Formal Expression
Adaptive capacity	Negative master functional	$\mathcal{A}(t) = -\mathcal{J}(t)$
Intelligence	Time-averaged rate of free energy reduction per unit novelty	$\mathcal{I} = -\langle (d\mathcal{J}/dt)/\mathcal{N}(t) \rangle_T$
Level-specific intelligence	Decomposed by RSP level	$\mathcal{I}_l = -\langle (dF_l/dt)/\mathcal{N}_l \rangle_T$
Recursive intelligence	Strange loop convergence efficiency	$\mathcal{I}_{\text{rec}}$: time-averaged rate of self-model convergence in social novelty
Meta-intelligence	Precision convergence speed	$\mathcal{I}_{\text{meta}}$: how fast $\Pi_l \rightarrow \Pi_l^*$ after change
Constitutive identity	Recursive intelligence $\leftrightarrow$ consciousness at Level 3b	$\langle \mathcal{I}_{\text{rec}} \rangle_T \geq 0 \iff \Phi(t)$ stable (given social novelty, RSP architecture)

The RSP theory of intelligence completes the framework's account of mind. At the recursive level, consciousness is not an epiphenomenal accompaniment to intelligence, nor intelligence a byproduct of consciousness. The recursive self-modeling architecture constitutes both: adaptive capacity in social environments (observed externally as intelligence) and phenomenal experience (undergone internally as consciousness). Total intelligence includes components at Levels 1 and 2 that are accompanied by their own grades of phenomenal character, each constituted by the computational process at that level. Valenced experience at Level 1. Structured experience at Level 2. Reflective self-awareness at Level 3b. The hard problem dissolves uniformly across all three levels by the same constitutive move: phenomenal character IS what the relevant process is like from within. In environments requiring recursive social prediction, intelligence is bounded without consciousness, and consciousness is idle without the adaptive pressures that sustain it. The master functional

$$\mathcal{I}$$

is the shared mathematical structure whose minimization simultaneously realizes both.

Index

A

Active inference — p.117 (App B) **Adaptive Resonance Theory (ART)** — compared p.68 (§X-C); complements RSP (resonance as implementation mechanism) **Affective-homeostatic schema** — p.10 (§III), p.19 (§IV), p.104 (App A), p.132 (App C) **Alien hand syndrome** — clinical dissociation p.62 (§IX, Prediction 6) **Anosognosia** — clinical dissociation p.62 (§IX, Prediction 6); comparison p.68 (§X) **Attention schema** — definition p.10 (§III); developmental role p.29 (§V), p.62 (§IX); in bat phenomenology p.45 (§VII); objections p.49 (§VIII); formal model p.104 (App A), p.117 (App B), p.132 (App C), p.147 (App D) **Attention Schema Theory (AST)** — compared p.68 (§X-B); extended by RSP (RSP adds 4 schemas + recursion + formalism) **Autism spectrum** — clinical dissociation p.62 (§IX, Prediction 6) **Autobiographical memory** — development p.29 (§V); predictions p.62 (§IX)

B

Barrett, Lisa Feldman — constructed emotion p.10 (§III); predictions p.62 (§IX); vs. basic emotions p.68 (§X) **Basic emotions theory** — compared p.68 (§X-G); vs. constructed emotion theory **Bennett, Michael Timothy** — Stack Theory compared p.68 (§X-I); w-maxing; Temporal Gap; convergence with RSP; p.207 (App H) **Biological universalism** — see Basic emotions theory **Body schema** — definition p.10 (§III); evolution p.29 (§V); bat example p.45 (§VII); objections p.49 (§VIII); predictions p.62 (§IX); formal model p.104 (App A), p.132 (App C)

C

Capgras delusion — clinical dissociation p.62 (§IX, Prediction 6) **Cascade control** — architecture p.147 (App D); developmental ordering p.29 (§V) **Cephalopod intelligence** — boundary test case p.219 (App H, Prediction H.4); high $\mathcal{I}_1 + \mathcal{I}_2$ without $\mathcal{I}_{\text{rec}}$ **Centrencephalic Proposal (CP)** — compared

p.68 (§X-E); complements RSP (CP as Level 1 substrate) **Constitutive identity proposition** — intelligence-consciousness identity at Level 3b, p.214 (App H §H.4); depends on §IV dissolution argument p.19 **Chalmers, David** — hard problem p.7 (§I), p.9 (§II), p.49 (§VIII) **Conceptual categorization** — Level 3a mechanism p.10 (§III); cross-cultural variation p.29 (§V), p.41 (§VI); in bat phenomenology p.45 (§VII); objections p.49 (§VIII); formal model p.104 (App A) **Consciousness State Space Model (CSS)** — compared p.68 (§X-D); extended by RSP **Consciousness, hard problem of** — introduced p.7 (§I); traditional framing p.9 (§II); dissolved p.19 (§IV), p.41 (§VI); reframed p.49 (§VIII); dissolved p.88 (§XI); free energy dissolution p.117 (App B) **Constructed emotion theory** — p.7 (Abstract); p.10 (§III, core framework); p.68 (§X, comparison); p.88 (§XI, conclusion) **Contraction mapping** — convergence proof p.147 (App D); referenced p.29 (§V) **Core affect** — definition p.10 (§III); as foundation p.7 (Abstract, §I); developmental emergence p.29 (§V); cross-cultural universality p.41 (§VI); bat phenomenology p.45 (§VII); objections p.49 (§VIII); predictions p.62 (§IX); comparison p.68 (§X); conclusion p.88 (§XI); formal model p.104 (App A), p.117 (App B) **Core schemas** — five schemas defined p.10 (§III); developmental role p.29 (§V); bat example p.45 (§VII); objections p.49 (§VIII); formal model p.104 (App A) **Cotard's delusion** — clinical dissociation p.62 (§IX, Prediction 6) **Cross-cultural variation** — as prediction p.41 (§VI); evidence p.62 (§IX); vs. basic emotions p.68 (§X) **Cross-Order Integration Theory (COI)** — compared p.68 (§X-B); extended by RSP

D

Damasio, Antonio — somatic markers p.10 (§III); proto-self p.10 (§III); cited p.104 (App A), p.117 (App B) **Default mode network** — strange loop substrate p.104 (App A) **Depersonalization** — social prediction failure p.19 (§IV); clinical dissociation p.62 (§IX, Prediction 6); control-theoretic model p.117 (App B), p.147 (App D) **Developmental progression** — 0–18 months through 7+ years p.29 (§V); mathematical necessity p.29 (§V); predictions p.62 (§IX); formal constraints p.117 (App B), p.132 (App C), p.147 (App D); intelligence-consciousness parallel p.207 (App H §H.6) **Dissociative identity disorder** — clinical dissociation p.62 (§IX, Prediction 6)

E

Emotion, four-fold characterization — surprise + gradient + evaluation + error; unified account p.166 (App E §E.6.5) **Emotional granularity** — defined p.10 (§III); interventions p.29 (§V); predictions p.41 (§VI), p.49 (§VIII), p.62 (§IX) **Epiphenomenalism** — distinguished from p.10 (§III) **Equivalence proofs** — prediction error = free energy gradient = TD error = control error, p.166 (App E §E.4) **Evolutionary origins** — valence to strange loops p.29 (§V); Level 1 deep time p.29 (§V); Level 2 mammalian p.29 (§V); Level 3 hominin p.29 (§V); Dunbar's number p.29 (§V) **Explanatory gap** — posed p.7 (§I), p.9 (§II); dissolved p.19 (§IV), p.41 (§VI), p.88 (§XI); free energy formulation p.117 (App B)

F

Feral children — evidence for RSP p.62 (§IX); Genie case p.29 (§V) **First-Order Representationalism (FOR)** — compared p.68 (§X-B); extended by RSP **Free energy, variational** — introduced p.7 (§I); developmental constraints p.29 (§V); predictions p.62 (§IX); conclusion p.88 (§XI);

full derivation p.117 (App B); relation to RL p.132 (App C); relation to control p.147 (App D) **Friston, Karl** — predictive processing p.10 (§III); free energy p.29 (§V); cited p.104 (App A), p.117 (App B) **Falsification conditions** — formalism-independent p.166 (App E §E.7.1); formalism-specific §E.7.2; nuclear option §E.7.3

G

Gauge symmetry (representational invariance) — internal coding arbitrary; p.166 (App E §E.6.1) **Global Neuronal Workspace (GNW)** — see Global workspace theory **Global workspace theory (GWT/GNW)** — compared p.29 (§V), p.68 (§X-A); complements RSP (RSP specifies broadcast content) **Graziano, Michael** — attention schema theory p.10 (§III); evolution p.19 (§IV); objections p.49 (§VIII) **Gurwitsch's Theory (GT)** — compared p.68 (§X-H); complements RSP

H

Higher-Order Thought Theory (HOT) — compared p.68 (§X-B); extended by RSP (RSP adds recursion + affect grounding) **Hofstadter, Douglas** — strange loop p.19 (§IV); objections p.49 (§VIII)

I

Information-theoretic parallel — data to self-referential computation p.29 (§V) **Integrated information theory (IIT)** — compared p.29 (§V), p.68 (§X-C); contradicts RSP (content-specific vs. content-neutral) **Intelligence, RSP theory of** — formal definition p.208 (App H §H.3); adaptive efficiency $\mathcal{I}$; level-specific decomposition; constitutive identity with consciousness at Level 3b p.214; predictions p.219 (§H.5) **Interoceptive monitoring** — Level 1 mechanism p.10 (§III), p.19 (§IV); formal model p.104 (App A), p.117 (App B)

M

Master functional ($\mathcal{J}$) — unified variational objective p.166 (App E §E.3); accuracy + temporal + complexity + recursive terms; all four formalisms as projections §E.3.2 **Master dynamics** — $\dot{\mathbf{x}}_l = -\kappa_l \partial \mathcal{J} / \partial \mathbf{x}_l$; simultaneously message passing, gradient descent, TD update, and control law; p.166 (App E §E.3.3) **Meaningful integration** — central thesis p.7 (Abstract, §I); defined p.10 (§III); mechanism p.19 (§IV); objections p.41 (§VI), p.49 (§VIII); bat example p.45 (§VII); predictions p.62 (§IX); comparison p.68 (§X); conclusion p.88 (§XI); formal model p.104 (App A), p.117 (App B) **Mirror self-recognition** — developmental milestone p.29 (§V); prediction p.62 (§IX) **Model-based planning** — Level 3 as planning p.132 (App C); strange loop as simulation p.19 (§IV)

N

Nagel, Thomas — bat question p.7 (Abstract, §I), p.9 (§II); misfires p.19 (§IV), p.45 (§VII); conclusion p.88 (§XI)

O

Orchestrated Objective Reduction (Orch OR) — compared p.68 (§X-F); contradicts RSP (classical vs. quantum)

P

Phantom limb — clinical dissociation p.62 (§IX, Prediction 6) **Phenomenal consciousness** — defined p.7 (Abstract); traditional view p.9 (§II); reconceived p.10 (§III) **Play, intrinsically motivated** — as exploration policy p.29 (§V); RL formulation p.132 (App C) **Precision as universal currency** — weights errors (A), balances accuracy/complexity (B), controls exploration ©, sets gain (D); unified account p.166 (App E §E.6.4) **Precision weighting** — attention as precision p.10 (§III); in prediction errors p.104 (App A), p.117 (App B); emotional regulation p.147 (App D) **Predictive processing / predictive coding** — framework p.7 (Abstract, §I), p.10 (§III); integration mechanism p.19 (§IV); formal model p.104 (App A), p.147 (App D) **Property dualism** — p.7 (§I) **Prosopagnosia** — clinical dissociation p.62 (§IX, Prediction 6) **Psychological Theory of Consciousness (PToC)** — compared p.68 (§X-D); extended by RSP

R

Quantum consciousness theories — compared p.68 (§X-F); contradicts RSP (classical vs. quantum) **Raw qualia** — critique p.7 (Abstract, §I), p.9 (§II); dissolved p.19 (§IV), p.41 (§VI); objections p.45 (§VII), p.49 (§VIII); conclusion p.88 (§XI) **Reaction to Reflection model (R2R)** — compared p.68 (§X-J); complements RSP; Value Saturation; perspectival entrapment; four evolutionary transitions (Trukovich, 2025) **Recurrent Processing Theory (RPT)** — compared p.68 (§X-E); complements RSP (RSP specifies recurrent content) **Recursive social prediction (RSP)** — introduced p.7 (§I); architecture p.10 (§III); mechanism p.19 (§IV); defended p.41 (§VI), p.49 (§VIII); applied p.45 (§VII); predictions p.62 (§IX); comparison p.68 (§X); conclusion p.88 (§XI); formal models p.104 (App A), p.117 (App B), p.132 (App C), p.147 (App D) **Reinforcement learning** — RSP as intrinsically motivated RL p.7 (§I), p.19 (§IV); full formulation p.132 (App C) **Romanian orphans** — evidence p.29 (§V), p.62 (§IX) **Russell, James** — core affect theory p.10 (§III); reward mapping p.132 (App C)

S

Sattin et al. (2021) — scoping review of 38 consciousness theories; basis for p.68 (§X) comparisons **Schizophrenia / thought insertion** — p.19 (§IV) **Self-model / self-schema** — definition p.10 (§III); strange loop p.19 (§IV); defended p.41 (§VI); bat p.45 (§VII); objections p.49 (§VIII); predictions p.62 (§IX); conclusion p.88 (§XI); formal model p.104 (App A), p.117 (App B), p.132 (App C), p.147 (App D) **Semantic Pointer Architecture (SPA)** — compared p.68 (§X-A); complements RSP **Social deprivation** — natural experiments p.29 (§V); predictions p.62 (§IX) **Social prediction** — mechanism p.10 (§III), p.19 (§IV); defended p.49 (§VIII); predictions p.62 (§IX); formal model p.104 (App A), p.117 (App B), p.132 (App C), p.147 (App D) **Spatial schema** — definition p.10 (§III); evolution p.29 (§V); bat example p.45 (§VII); objections p.49 (§VIII); predictions p.62 (§IX); formal model p.104 (App A), p.132

(App C) **Stack Theory (Bennett, 2025)** — compared p.68 (§X-I); complements RSP; w-maxing, tapestries of valence, Temporal Gap **Strange loop** — defined p.7 (§I); mechanism p.19 (§IV); objections p.49 (§VIII); convergence proof p.147 (App D); free energy p.117 (App B); RL formulation p.132 (App C); predictive coding p.104 (App A)

T

Theory of mind — developmental role p.29 (§V); objections p.49 (§VIII); predictions p.62 (§IX); comparison p.68 (§X); RL formulation p.132 (App C) **Three-level necessity theorem** — all four formalisms independently require three levels; p.166 (App E §E.6.3) **Timescale separation** — cascade design p.147 (App D); developmental constraints p.117 (App B), p.147 (App D); referenced p.29 (§V); unified falsification p.166 (App E §E.7.1) **Transfer functions** — Level-specific dynamics p.147 (App D) **Trukovich, Joseph J.** — R2R model compared p.68 (§X-J); Value Saturation; perspectival entrapment; independent convergence with RSP

V

Value Saturation (Trukovich) — consciousness as explicit recursive self-modeling saturated by homeostatic significance; compared p.68 (§X-J) **Unified mathematical framework** — master functional, equivalence proofs, emergent properties; p.166 (App E) **Valence / arousal** — core affect p.7 (§I), p.10 (§III); reward signal p.19 (§IV); cross-cultural p.41 (§VI); bat p.45 (§VII); objections p.49 (§VIII); predictions p.62 (§IX); conclusion p.88 (§XI); formal model p.104 (App A), p.117 (App B), p.132 (App C); free energy–value duality p.166 (App E §E.4)

W

W-maxing (Bennett) — generalization-based intelligence; compared p.68 (§X-I); conjectured equivalence with RSP's free energy minimization p.208 (App H §H.3.6) **Weakness hierarchy** — Level 1 strongest/most specific → Level 3 weakest/most general; p.208 (App H §H.3.6)
