

Daniel Simon Jr.

Beyond the Bat: Dissolving the Explanatory Gap Through Meaningful Integration — I. The Dissolution

by Daniel Simon Jr.

Abstract

The "hard problem" of consciousness rests on an assumption that has gone largely unexamined: that phenomenal experience consists of raw qualia, pre-conceptual subjective states existing independently of meaning. This paper argues that the assumption is false, and that recognizing this dissolves rather than merely narrows the explanatory gap.

I propose the Recursive Social Prediction (RSP) model, a three-level architecture in which phenomenal consciousness is constituted by 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 structures: 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 turns inward to model the self modeling others modeling the self. The central claim is constitutive identity: phenomenology *is* recursive self-modeling, not a separate property that could be absent from it. The explanatory gap is not bridged but dissolved.

The paper defends this dissolution against the major philosophical objections — Jackson's knowledge argument, the zombie argument, Descartes' Cogito, and the epiphenomenalism charge — and reframes Nagel's bat question: not "What is it like?" but "What meaningful integrations does a bat's architecture perform?" Empirical evidence and mathematical formalizations supporting the architecture are presented in companion papers (Parts II and III of this series).

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

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 products of the brain's integration of biological core affect (valence and arousal) with culturally learned conceptual categories through recursive self-modeling. I propose a specific architecture for this integration: the Recursive Social Prediction (RSP) model, a three-level hierarchy formalized in a companion paper (Part III of this series) across four mathematical traditions — hierarchical predictive coding, variational free energy, intrinsically motivated reinforcement learning, and cascade control theory — unified by a single master functional in Appendix E of that paper.

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 does not exist. The questions need reframing, not answering.

Why This Matters

This is not mere conceptual housekeeping. The raw qualia assumption 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. And it prevents productive dialogue between consciousness research and emotion theory, developmental psychology, and cultural neuroscience.

If the argument here succeeds, the result is not just conceptual clarity but new empirical research programs and productive connections across previously isolated domains. The empirical evidence supporting the RSP architecture, including evolutionary, developmental, and clinical evidence along with seven falsifiable predictions, is presented in Part II of this series. The mathematical foundations, including a conditional contraction-mapping theorem for the strange-loop dynamics (Part III, Proposition E.4) and the algorithmic specification of the unified architecture (Part III, Algorithm 1), are in Part III.

Previous attempts to dissolve the hard problem, from Dennett's (1988) 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 self-aware phenomenal consciousness, drawing on Chalmers's (1996) framing of the hard problem and convergent insights from phenomenological and analytic traditions. The result is a theory that is simultaneously empirically grounded (Part II), mathematically formalized (Part III), and philosophically defended as structurally necessary.

Structure

Section II reviews the traditional framework and its dependence on the raw qualia assumption. Section III introduces the RSP architecture and the meaningful integration thesis. Section IV presents the core dissolution argument: why phenomenology *is* recursive self-modeling. Section V reframes Nagel's bat question as a tractable empirical program. Section VI addresses twelve philosophical objections — the inverted spectrum, other minds, the explanatory gap, cultural variation, introspective access, the "why feel anything" challenge, the functionalism charge, Jackson's knowledge argument, Descartes' Cogito, the epiphenomenalism charge, the distinctive realness of ordinary experience, and unconscious prediction under anaesthesia. Section VII concludes.

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 irreducibly 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 (Russell, 1912; Ayer, 1940) went further, explicitly positing raw sense-data as the foundation of perceptual knowledge.

Contemporary philosophy retains this framework. Chalmers' (1995) "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

A note on scope. The architecture introduced here operates at three characteristic timescales (sub-second to early-second for Level 1 core affect, ~1–10 s for Level 2 schema integration, ~10 s to minutes for Level 3 recursive social prediction). Specific neural-substrate claims and mathematical specifications — including the evidence linking each level to anterior insula and anterior cingulate cortex (Level 1), default mode network and temporoparietal junction (Levels 2–3), and visceromotor and self-reflection circuits — are developed in Part II §§3–5. The present paper retains the philosophical-architectural argument, with anatomical claims kept to the minimum the dissolution thesis requires.

Core Affect as Foundation

I begin with something genuinely pre-conceptual: **core affect** (Russell, 1980, 2003; Barrett & Bliss-Moreau, 2009) — neurophysiological states characterized by two dimensions, **valence** (pleasant/unpleasant) and **arousal** (activated/calm), Russell's affective circumplex. These states arise from interoceptive monitoring of bodily conditions and subcortical processing. Their

mammalian elaboration is well documented (Panksepp, 2005), with valenced approach/avoidance systems broadly distributed across vertebrates (deeper bilaterian roots in Part II). Core affect operates **continuously** as a background state; remains **undifferentiated**, not yet carved into specific emotions; yet is **functionally significant**, connecting directly to approach/avoidance and homeostatic regulation.

Core affect is not yet experience in the *self-aware* sense the hard problem targets, but it carries genuine phenomenal valence. An infant in unpleasant high-arousal core affect does not yet feel "angry" or "scared"; it feels *something* global and undifferentiated. Core affect is the phenomenal ground of feeling, not yet conscious emotional experience.

Core Schemas: The Bridge Between Biology and Concepts

Between core affect and high-level categorization lies an intermediate structural layer: **core schemas**, pre-reflective, continuously operating internal models that provide the foundational architecture for conscious experience (Gallagher, 2005; Metzinger, 2003; Graziano, 2013).

Research converges on five fundamental schema types. **Body schema** comprises pre-reflective sensorimotor capacities forming an implicit spatial-proprioceptive framework, operating below awareness to structure how you experience your body's location, posture, and capabilities — when you reach for a cup, you don't calculate limb positions; the body schema handles it transparently.

Self-schema creates the phenomenal self-model generating the unified sense of being a coherent entity persisting through time: mineness (this experience is *mine*), perspectivalness (I am the center of experience), and continuity (the same "I" across time). The **attention schema** is the brain's simplified model of its own attention (Graziano, 2013); it evolved for social cognition (modeling where others direct attention), and consciousness is the same machinery turned inward — no separate process added. The regions computing "Person X is aware of Y" also compute "I am aware of Y."

Spatial schema provides frameworks structuring not just literal space but abstract cognition. Hippocampal-parietal systems scaffold the organization of memory, conceptual knowledge, temporal relations, and social understanding; mental models are inherently spatial (memories are "places," time flows "forward," hierarchies "above" or "below"). The **affective-homeostatic schema** maintains continuous interoceptive monitoring that *constitutes* primordial feeling — Damasio's (1999) "proto-self," the elementary feeling of existence.

These schemas operate pre-reflectively, exhibit transparency (experienced as reality rather than representations), perform control functions, and structure higher-level processing. **They are**

not learned concepts but developmental achievements. Infants don't learn to have a body schema; it develops through sensorimotor experience. The self-schema emerges in the toddler period, with mirror self-recognition typically reported at 18–24 months in WEIRD samples (Amsterdam, 1972; Lewis & Brooks-Gunn, 1979), though recognition rates vary dramatically across non-Western populations — some near zero — so the milestone may reflect culturally specific modes of self-expression rather than a universal developmental threshold (Broesch et al., 2011). These are proto-concepts: structural organizations that enable conceptual thought rather than products of it.

Conceptual Categorization: From Schemas to Meanings

Building on these 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 feeling. When your brain categorizes current core affect (unpleasant, high-arousal) as an instance of "anger" (given context suggesting provocation), feeling angry — its particular quality, action tendencies, and social meanings — is constituted by that categorization.

But categorization operates *through* the core schemas. The concept "anger" integrates with body schema (clenched muscles, forward-leaning posture), self-schema (I am the wronged party), attention schema (focused on the offender), spatial schema (the offender is "against" me), and affective schema (unpleasant high-arousal with approach motivation). The feeling of anger is not the concept applied to core affect; it is the concept integrated through all these schemas at once.

Several converging lines support this constitutive role. **Cross-cultural variation:** emotions don't translate cleanly because cultures have genuinely different emotion concepts creating 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 experiences, not just better descriptions of identical ones (Barrett et al., 2001; Kashdan et al., 2015). **Developmental construction:** children don't gradually access pre-existing emotions; they acquire differentiated experiences as they learn emotion concepts (Widen & Russell, 2008).

Predictive Integration

Predictive processing frameworks (Clark, 2013, 2016; Friston, 2010; Seth, 2013) add the mechanistic piece: the brain is a prediction machine, and perception is not passive reception of

sensory input but active prediction of its causes.

When you see something red, your brain doesn't receive raw red-data and then categorize it. It predicts "This is RED" from prior learning, and this prediction shapes early sensory processing through top-down connections, integrates with emotional associations (red-danger, red-passion), activates semantic knowledge, generates expectations, modulates attention and arousal, and prepares actions. The *experience* of seeing red is this integrated prediction; there is no stage at which "raw redness" exists independently of it — the what-it's-like of red is the brain's multi-layered prediction integrating sensory evidence with conceptual knowledge, emotional associations, and contextual expectations.

Meaningful Integration: The Full Picture

The full picture: **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.** The three-level architecture:

Level 1: Core Affect (biological foundation)

Valence (pleasant/unpleasant) and arousal (activated/calm); subcortical, continuous, undifferentiated.

Broadly distributed across vertebrates (deeper bilaterian roots in Part II).

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.

Constitute structured phenomenality in animals and pre-verbal infants.

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

Level 3a — Conceptual categorization: Emotion categories, semantic knowledge, linguistic labels — learned socially, culturally variable, enabling 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 component, detailed in the next section, *constitutes self-aware phenomenality* — the "for

me" character of mature human experience that the hard problem targets. It does not constitute phenomenality as such; the constitutive identity claim applies specifically to the recursive, self-aware dimension, while simpler phenomenality exists at Levels 1 and 2 (§VI, "Why Level 3b, Not Level 2?").

Implicit precursors are reported in great apes (Krupenye et al., 2016; conceptually replicated in Kano et al., 2019), although the great-ape mindreading literature shows mixed results across paradigms and labs (Krupenye & Call, 2019, for review), and the comparative literature on non-human theory of mind is contested (Heyes, 2015; Penn & Povinelli, 2007). Full linguistic elaboration and explicit recursive depth appear specific to humans.

To restate the key hedge: phenomenality here is graduated rather than threshold. Lower levels carry genuine, simpler phenomenal character; the recursive, self-aware dimension the hard problem targets is what Level 3b distinctively constitutes (full argument in §VI, "Why Level 3b, Not Level 2?").

Full human consciousness exists at no single level. It is not in raw sensory transduction (meaning-free), nor core affect alone (undifferentiated valence), nor schemas alone (structural, not specific), nor abstract concepts alone (disembodied), nor predictions alone (possibly unconscious). Their integration, experienced from within, *is* the feeling. When 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 qualitative character.

[Figure 1. The RSP Three-Level Architecture — figure omitted for blind review]

Why This Three-Level Architecture Matters

This layered view resolves several puzzles. **Animal consciousness without language** becomes comprehensible: animals have core affect and core schemas but lack linguistic concepts, giving them genuine structured experience — not "raw feels" but schema-organized affects. A dog experiences fear (unpleasant high-arousal organized through body, attention, and rudimentary self-schemas) without the concept "fear."

Infant development follows a clear arc: from core affect with rudimentary schemas (birth), through increasingly schema-structured experience (0–18 months), to conceptually elaborated feelings (18+ months) — each stage genuinely conscious but increasingly sophisticated. **Cross-cultural variation with biological universals** makes sense: all humans share core affect and develop similar schemas, but cultural concepts vary — explaining both universality and diversity.

The role of social cognition follows from the attention schema's inherently social nature: Graziano's (2013) theory proposes that consciousness is the brain's model of attention, evolved for social purposes — making consciousness social-perceptual from its schematic foundation.

Causal Indispensability of the Organizational Level

The dissolution defended here does not reduce consciousness to underlying processes while making phenomenology causally inert; meaningful integration has causal efficacy at its own level of description. The experience of fear causes flight — not merely the underlying neural activity, but the integrated meaningful state (fear) realized at the architectural level. The phenomenology shapes subsequent predictions, attention, learning, and action; the feeling is an active component of the mind's predictive architecture, not an inert by-product.

The full response to Kim's (1998, 2005) causal exclusion argument is developed as Objection 10 in §VI. In summary, the RSP model commits to non-reductive physicalism: consciousness is constituted by physical processes organized in a specific way, and the organizational description is causally indispensable in the same way thermodynamic regularities are indispensable to physical explanation.

The Transcendental Structure of Consciousness

The three-level architecture is more than one way consciousness happens to be organized in mammals: it represents the *conditions of possibility* for the kind of consciousness we can coherently describe and investigate. The parallel with Kant (1781/1998) is instructive. Kant argued that space, time, and causality are not features of a world we happen to encounter but preconditions for any 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 system. A system for which nothing is better or worse has no stake in its processing, no vantage point from which experience could be *for* anything. Valence is the minimum condition for a subject position; without it, 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 organization is, in James's (1890) phrase, a "blooming, buzzing confusion." For experience to be *of* something — to have content and not merely tone — it must be organized: body schema furnishes the frame in which content is located, spatial schema the relational structure between its parts, attention schema makes some content figure rather than ground, and affective-homeostatic schema connects content to

regulatory imperatives. Without structure, valence has nowhere to attach. Level 2 is the minimum condition for content-bearing experience.

Level 3 is necessary because content-bearing experience without a recursive self-model is not yet self-aware experience. The consciousness the hard problem targets — the "what it is like *for me*" character — requires that the system model itself as the one undergoing the experience, with that self-model available to its own predictive operations. A creature with Level 1 valence and Level 2 schemas has structured felt experience; what it lacks is the recursive standpoint from which experience is registered *as belonging to itself as an experiencer*. Recursive self-modeling — the machinery used to model others' minds, applied reflexively to the system's own predictions — introduces that "for me" character. The argument is transcendental, not empirical: any system whose phenomenology has the perspectival, self-attributing character of the hard problem's explanandum must implement recursive self-modeling, since there is no other way to constitute it. Level 3 is the minimum condition for self-aware phenomenality. Together, Levels 1–3 give the conditions of possibility for the kind of consciousness the explanatory gap was supposed to be a gap *about*.

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 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 must 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 that self-model is used predictively in social contexts.

The Strange Loop Reveals Its Purpose

Hofstadter (1979) identified consciousness as a "strange loop," a recursive, self-referential structure in which consciousness observes itself observing itself. But he left the functional purpose of this recursion 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 self-admiration 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 a prediction machine (Friston, 2010; Clark, 2016), and to predict its own future states it needs a model of itself: its current state, dynamics, tendencies, and patterns. The way you can observe yourself thinking is not philosophical curiosity but functional necessity. Consciousness is the continuously updated self-model that enables prediction of future mental states.

[Figure 2. The Strange Loop: Recursive Social Prediction — figure omitted for blind review]

Concrete Examples of Predictive Self-Modeling

Consider deciding what to say in a conversation. You don't speak randomly: you simulate possible utterances, predict how you would feel saying each and how the other person would respond, and select the option whose predicted consequences are most favorable. This requires simulating your own future mental states — 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 (a Schneiderian first-rank symptom in schizophrenia): Patients experience thoughts as alien — inserted by external forces. On the comparator model of agency (Frith, 1992), the self-model that should register "I generated this thought" fails to update on its own efferent activity. Phenomenology changes precisely because the predictive self-model is impaired.

Ventromedial prefrontal damage: Patients lose the capacity to anticipate the affective consequences of their actions and show characteristic decision-making deficits (Bechara et al., 1997). They can't predict "if I say this, I'll feel regret later" because the self-model enabling future affective prediction is damaged.

Depersonalization-derealization disorder: Patients feel detached from their own mental processes, as if observing themselves from outside (Sierra & David, 2011; Medford et al., 2016). The self-model still operates but is not properly integrated with first-person experience.

These clinical cases show that predictive self-modeling is a concrete cognitive function: when impaired, consciousness changes in specific, measurable ways.

Consciousness as Relational and Social

The crucial point: **you model yourself primarily in relation to others**. The self-model is not a solipsistic internal representation but is inherently relational, structured around social contexts and predictions. So much of conscious experience involves modeling others — predicting what someone is thinking, anticipating reactions to your words, simulating conversations before having them. Even in solitude, thinking remains implicitly relational: "I should exercise today" models your future self and how it will feel; remembering an embarrassing moment requires modeling how others perceived you; planning for a meeting requires modeling attendees' responses. Consciousness continuously models 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, and cooperation, competition, alliance-formation, deception detection, and coordinated group activity have plausibly all shaped human cognitive evolution. The social-brain hypothesis (Dunbar, 1998) proposed that primate neocortex size scales with social group size; more recent work shows that ecological factors — particularly diet — predict primate brain size at least as well as sociality (DeCasien et al., 2017; Powell et al., 2017). The framework defended here does not require the strong claim that social pressures uniquely drove recursive self-modeling. The weaker claim suffices: whatever the selective pressures, recursive social prediction is one architectural pattern supporting both other-modeling and self-modeling, and this pattern, when running, constitutes the self-aware phenomenology distinctive of mature human experience.

The architectural logic turns on this: to predict how others will respond, the system must predict what they take to be the case about it; to do that, it must model itself as they would model it. This recursion — modeling oneself using the machinery used to model others — produces a perspectival self-model that can be viewed from others' standpoints.

The neural machinery supporting theory of mind is plausibly recruited for self-modeling. Graziano's (2013) Attention Schema Theory makes this explicit: consciousness is the brain's mechanism for modeling others' attention applied to its own — constitutive identity, not causal production. Two related but partially distinct cortical networks have been implicated: the mental-state-attribution network (right temporoparietal junction, posterior superior temporal sulcus), which tracks the contents of others' minds (Saxe & Kanwisher, 2003; Schurz et al., 2014), and the self-reflection network (medial prefrontal cortex, precuneus, posterior cingulate cortex), which tracks one's own. The architectural prediction is that recursive self-modeling

recruits both, co-deploying the self/other-tracking machinery when the system models itself as one of the social agents the network represents. The precise overlap and conditions of joint engagement remain under investigation, and the claim is stronger than any single result establishes; the empirical bridge is developed in Part II.

This is why consciousness is inherently social-architectural, built from other-modeling machinery: perspectival (models are always from a viewpoint — one's viewpoint imagining others'), predictive (it exists to predict behavior, especially social behavior), empathically structured (it uses the same systems for self and other), and recursive (it models the self modeling others modeling the self).

Why This Explains the Feeling of Consciousness

We can now 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 generates core affect, organizes it through schemas, categorizes it with concepts, predicts your own future mental states, models yourself as others might perceive you, and feeds these predictions back into itself, the integration, prediction, and recursion *constitute* the feeling rather than producing it as a separate by-product. Phenomenology is what the process is like from inside the system — just as visual experience is not something added to visual processing but what that processing is like from inside.

This also explains why consciousness is continuous and self-sustaining. The predictive loop maintains itself: at moment t , your state generates predictions about $t+1$; those predictions shape what happens; the prediction error updates the self-model, which generates new predictions for $t+2$ — the cycle that creates the continuous stream of consciousness. And the predictions concern not just what you'll think but how your thoughts will affect others and how their responses will affect you ("If I think X, I might say Y, they might respond Z, I would then feel W...") — social simulations requiring recursive self-modeling in social space.

Dissolving the Explanatory Gap Through Social Prediction

Return now to the explanatory gap. The traditional framing asks why physical processes (neurons firing) should *create* separate phenomenal properties (raw qualia); so framed, the gap appears unbridgeable. The meaningful-integration framing improves matters — physical processes *are* meaningful integrations of affect, schemas, and concepts, viewed at the architectural level — but still leaves open why such integration would have phenomenal

character. The social-predictive framing closes the question: the running of recursive self-modeling for social prediction is phenomenology.

Two recent, independent lines of work illustrate the same dissolution move applied to adjacent phenomena. Barenholtz's (2025) analysis of large language models argues that linguistic meaning is not a representation of something beyond the generative process but the operation of relational structure generating its own continuations — structurally the same replacement of "grounded in" with "is the operation of" that RSP performs for phenomenal consciousness (Barenholtz's essay concerns language, not consciousness, and is not peer-reviewed; the parallel is structural, not evidential). And Singleton and Fotopoulou's (2026) finding that the sense of "being real" is mechanistically continuous across virtual, dream, and waking contexts supports the corollary that felt realness is not a metaphysical surplus but the operation of the predictive self-model — a point Paper II develops on the clinical side.

The constitutive-identity claim should be read as a Type-B a posteriori identity in the Kripkean sense (Kripke, 1980): phenomenal properties and architectural-functional properties are picked out by different concepts (phenomenal concepts; architectural-functional concepts) but refer to the same fact. This is the same logical form as "water is H₂O" or "temperature is mean kinetic energy" — necessary identity, knowable only a posteriori, with the appearance of a residual gap arising from the distinct concepts rather than from any further metaphysical layer. Phenomenal concepts have no independent primary intension that would let one coherently conceive of the architecture without the phenomenology, because phenomenal concepts are themselves constituted by the architecture's recursive operation on its own states (a constraint developed in §VI Objection 6's reply to the zombie argument). On this reading, "the architecture is described from outside, the phenomenology is registered from within" is a claim about two modes of access to one fact — not, as a Fregean reading would have it, a claim that there are two facts standing in a sense-reference relation. There is one fact. The dissolution is the recognition that the apparent gap was an artifact of holding the two concepts apart while looking for a bridge between them.

The gap dissolves because every aspect of the phenomenon becomes tractable. The functional role is clear: recursive self-modeling enables social prediction, evolutionarily essential for social species. The mechanism is specifiable: identifiable neural structures (TPJ, mPFC, posterior cingulate), the schemas they implement (attention schema, self-schema), and the recursive loops they create. Development is traceable through social interaction; phylogenetic gradation exists (consciousness scales with social complexity and recursive capacity rather than being all-or-nothing); and impairment is predictable, since 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" — presupposes that phenomenology could be present or absent while leaving the functional architecture intact. The transcendental argument of §III reinforces why it cannot. 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 should *produce* experience rests on a false premise. It assumes the architecture could exist without the experience; but if the architecture constitutively *is* the experience, there is no gap to bridge.

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. Likewise, phenomenology is not something extra added to recursive self-modeling; it is intrinsic to the recursive structure experienced from within. Remove it and you no longer have consciousness — you have unconscious self-monitoring.

This is why purely philosophical zombies (beings physically identical to us but lacking phenomenology) are incoherent. You can't have the structure without the experience any more than you can have a running program without computation occurring.

The Dissolution in Philosophical Context

The dissolution strategy pursued here invites comparison with several established positions in the philosophy of consciousness.

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 does not claim consciousness is illusory. Consciousness is *real* and constituted by recursive self-modeling; what is illusory is only the *characterization* of it as involving non-functional intrinsic properties. Phenomenal properties exist, but as functional properties of the recursive architecture, not substrate-intrinsic qualities that float free of organizational structure.

Access consciousness and phenomenal consciousness. Block (1995, 2007) distinguished access consciousness (information poised for reasoning, reporting, and behavioral control) from phenomenal consciousness (the "what-it-is-like" character that, he argues, can overflow cognitive access). RSP rejects the distinction. All consciousness is constituted by recursive self-modeling, which is itself a form of cognitive access: the recursive self-model's access to its own ongoing predictive operations. The apparent overflow Block reports — that

visual experience contains more than is reportable — is, on the RSP account, the difference between (a) the rich Level 1 + Level 2 schema-organized representations the recursive self-model integrates over, and (b) the comparatively narrow Level 3a categorizations encoded into discursive report. Both are forms of access; neither is access-free phenomenology. The "overflow" is the gap between schema-level integration and conceptual-level reporting, not evidence for phenomenology outside any access relation. This does not deny the phenomenology Block points to but re-describes it as a different mode of access — the recursive system's access to its own running — rather than a residue left over after access is subtracted.

V. 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: it presupposes raw, meaning-independent phenomenology accessible if only we had the right vantage point. The 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.

Nagel's Deeper Argument: The Subjective-Standpoint Asymmetry

Nagel's (1974) argument is sometimes read as a claim about raw qualia that the dissolution above answers. But Nagel's deeper argument is structural: scientific objectivity proceeds by abstracting away from any particular standpoint, while consciousness is essentially the having of a standpoint, so scientific description is constitutively inadequate to capture conscious character. A complete physical description of a bat could in principle be given from the third-person standpoint, but the first-person character of being a bat — the world disclosed from this perspectival origin — is what such a description leaves out.

The RSP response turns on a single observation. The first-person standpoint Nagel takes to be ineliminable from consciousness is, on the constitutive-identity claim, the recursive self-model running. It is not a residue left over after objective description finishes; it is what the recursive operation *is*, viewed from within itself. Nagel is right that the standpoint cannot be captured by a description that abstracts away from any particular standpoint, and right that it is essentially perspectival. He is wrong only in the further inference that the perspectival character must therefore be a non-physical residue. On the architectural-functionalist alternative, the perspectival character is what a recursive self-modeling architecture *is* when running. The asymmetry between objective description and subjective standpoint is real—it is the asymmetry

between (a) describing a recursive system from outside and (b) being that system's running—but neither side eliminates or is non-physical relative to the other.

This respects Nagel's insight without conceding the dualistic conclusion. The objective/subjective asymmetry is preserved; what is denied is that it indicates two ontologies. There is one architecture, describable in the third person and constitutive of a first person when it runs.

What To Ask Instead

Instead of "What is it like to be a bat?" I propose three types of question that yield empirical traction.

Architectural questions focus on functional organization. Approach/avoidance behaviors, distress calls, and pleasure responses reveal what core affects bats experience; discrimination learning reveals their conceptual categories; navigation and prey-anticipation behaviors reveal the sophistication of their predictive models; episodic-like memory and delayed gratification reveal the scope of temporal integration; recognition, social learning, and cooperation studies reveal the self-other models they construct.

Ecological questions examine how environmental pressures shape consciousness. How does echolocation shape bat predictive models differently from vision, and what affordances does each provide that the other does not? How do bat social structures shape the categories bats learn, and how did evolutionary pressures push bat consciousness in directions different from primates?

Comparative questions situate bat consciousness phylogenetically. How does it compare to dolphin consciousness (both echolocate) or bird consciousness (both navigate 3D space)? Which aspects are convergently evolved across mammals, and which are lineage- or niche-specific?

What We Can and Cannot Know

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

We can know several things with confidence. Approach/avoidance behaviors, distress vocalizations, and pleasure responses indicate that bats have core affect. We can map their core schemas: a **body schema** highly sophisticated for flight and 3D navigation; a **spatial schema** building allocentric cognitive maps for roosts and foraging routes; an **attention schema** modeling where conspecifics direct attention, essential for social roosting; a **self-schema** likely present but simpler than primate self-models; and an **affective-homeostatic schema** monitoring hunger, fatigue, and threat. These schemas differ structurally from human ones in

predictable ways: the body schema integrates echolocation (temporal/auditory) rather than vision, the spatial schema emphasizes volumetric over surface space, and the attention schema tracks conspecifics but likely lacks theory-of-mind sophistication. Bat consciousness is genuine structured experience organized through these schemas.

We cannot know certain things, and we don't need to. What raw bat-echolocation-qualia "feel like" is meaningless because no such thing exists. Whether bat-red matches human-red is unanswerable because concepts and schemas don't map this way. Whether bats have "true" versus mere "as-if" consciousness is a false dichotomy built on raw-qualia assumptions.

The key insight: bat consciousness differs from human consciousness not in having different raw qualia but in having differently structured schemas—organized around echolocation and 3D flight—combined with simpler conceptual elaboration (fewer learned categories, no linguistic concepts). These differences are empirically investigable: behavioral discriminations reveal conceptual categories, neural organization reveals schematic structure, generated predictions reveal forward models, and responses to novelty, threat, and reward reveal affective organization.

The question dissolves because it rests on false assumptions. "What is it like to be a bat?" presupposes a determinate fact about raw bat-phenomenology accessible in principle. But phenomenology is constituted by meaningful integration through schemas and concepts, and bat integrations differ from human ones at both levels. There is no neutral standpoint from which to compare raw feels, because raw feels do not 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 is constituted by integration across multiple levels.

Biological processing occurs as the bat's nervous system analyzes echo delays, Doppler shifts, and amplitude modulations, producing activation across auditory cortex, superior colliculus, and association areas. **Core affect** is the system's valenced state when the object is so registered—pleasant low-arousal for a safe roost-mate, unpleasant high-arousal for a predator. The valence is not a separate output of the biological processing; it is what that processing constitutes at the architectural level of valence and arousal.

Schema integration constitutes what bat consciousness distinctively *is*, across several schemas at once. Through the **body schema**, the proprioceptive model fuses flight mechanics with echo feedback—where the body sits relative to the object, what wing adjustments are needed, whether distance falls within capture range—not as cognitive calculation but as pre-reflective sensorimotor organization. The **spatial schema** organizes objects not as isolated entities but as locations in navigable 3D space ("roost entrance is above and behind me," "moth moving left-forward at closing speed"). The **attention schema** maintains an internal model of attention directed at the moth—not the thought "I am aware of the moth"—and tracks where roost-mates direct theirs, enabling social coordination. The **self-schema** supplies the sense of a subject to whom these experiences belong and an agent who will act: when the bat recalls catching similar prey yesterday, the same "I" carries the continuity. The **affective-homeostatic schema** monitors hunger, fatigue, and satiation, supplying the motivational context—why this moth matters right now.

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

Prediction integrates as the bat anticipates the object's path, prepares motor commands, and forecasts the sensory consequences of action: "moth, moving left, will be at location X in 100ms; prepare capture maneuver; expect taste/satiation upon success."

Phenomenology is this integrated prediction—structured through schemas, categorized by learned concepts, viewed from inside the bat's running architecture. Not raw echo-feels, but meaningful integrated structure: "prey-to-catch-now," organized through spatial navigation, body positioning, attention, 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. **Structurally**: the body schema is built around flight-and-echolocation rather than bipedal-and-vision integration; the spatial schema emphasizes volumetric 3D over surface-based 2D space; the attention schema involves simpler social modeling than sophisticated theory-of-mind; the self-schema is present-focused, with limited episodic memory rather than extended autobiographical narrative. **Conceptually**: bats hold simpler learned categories (dozens to hundreds versus thousands), with no linguistic elaboration to create fine-grained distinctions and less abstraction from immediate sensorimotor contexts.

We can understand these differences functionally without accessing raw bat-feels. "What is echolocation like?" becomes "How does echolocation structure bat schemas differently from how vision structures human schemas?"—and that is empirically tractable through neural organization, behavioral discriminations, navigation and hunting, and social behavior. The mystery is not metaphysical (ineffable raw qualia) but empirical (structural organization we are still mapping).

VI. Objections and Replies

The RSP architecture faces several philosophical challenges. Each dissolves under the meaningful integration account.

Objection 1: The Inverted Spectrum and Privacy of Qualia

The traditional problem in its sharpest form: if qualia are raw, private, intrinsic properties, how could we ever know whether your red matches my red? Perhaps your color experiences are systematically inverted relative to mine—what you call red, I would experience as green—but since we both learned to call the long-wavelength color "red," we never detect the inversion.

Under meaningful integration, this dissolves. There is no "raw red experience" to compare. Instead there is **biological similarity**—our visual systems share structure, so red surfaces produce similar activation and similar core affect (arousal, attentional salience). There is the **cultural concept** RED, learned from shared language, carrying semantic associations (blood, roses, stop signs), valences (danger, passion), and functional roles largely held in common. And there is **personal history**—your specific encounters build a unique associative network (your grandmother's red coat, going red with embarrassment). **Integrated phenomenology** weaves biological response, concept, and personal association together. So does mine. Our experiences are neither identical (different histories) nor private (shared biology and culture): they are **intersubjectively similar**—enough to ground communication, different enough to remain unique. We communicate about red not despite but *because* the experience is meaningfully integrated; meaning provides the overlap.

The inversion scenario then becomes incoherent at two levels. The **weaker version**—partial inversion, where our reds differ phenomenally while we use the same word for the same wavelengths—is answered by the semantic network. Concept learning depends on pattern-matching across modalities: emotional valence and downstream associations with blood, danger, heat. Your experience of red is partially constituted by engagement with this network; if your long-wavelength experience genuinely failed to integrate with its other nodes, you would

fail to learn the concept correctly. The scenario thus stipulates a phenomenal difference that requires a corresponding integrative difference, and its supposed undetectability collapses.

The **stronger version** stipulates total functional inversion: the integrative network itself is inverted, leaving no behavioral or functional trace. This presupposes a substrate-neutral phenomenal property, attached to functional structure as a separable layer, that could be inverted while the structure is held fixed. The constitutive-identity claim denies precisely this: there is no phenomenal layer separable from the integrative architecture, so nothing is left over to invert. The total-inversion scenario does not pick out a coherent metaphysical possibility—it posits a difference where, by the constitutive thesis, none can exist.

Objection 2: The Problem of Other Minds

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 inner experience; perhaps animals lack consciousness entirely.

This assumes consciousness is all-or-nothing, an inner light either present or absent, verifiable only in my own case. Under meaningful integration, consciousness is a matter of degree and kind, and the relevant questions become empirical rather than metaphysical:

- 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 find answers through behavior, neuroscience, and comparative psychology. We need not telepathic access to another's raw qualia but an understanding of their functional architecture.

For other humans: shared biology, culture, and language give strong reason to believe their integrations resemble ours. They learn the same emotion concepts and respond to similar situations with similar patterns. Their consciousness is not mysterious but 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) and likely 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 different raw qualia but by different integrations suited to different ecological niches.

For AI systems: artificial systems are diagnosed by architecture, not substrate. This paper's architectural-functionalist commitment is that consciousness is constituted by a specific recursive self-modeling architecture, not by biological tissue per se; whether contemporary AI instantiates that architecture is a question about structure, not silicon. The full diagnosis of large language models, agentic systems, and embodied foundation models—and the dispatch of Searle's Chinese Room—is developed in §VI Objection 7 below, where the architectural criteria are stated and applied. The summary verdict: current AI satisfies the criteria only degenerately, and the position is falsifiable, specifying what structural changes would warrant revising the prediction.

Objection 3: The Explanatory Gap

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 floating free of functional role, so that no functional description explains why a state should have a particular raw feel. But if experiences are meaningful integrations, the gap dissolves. The task is no longer to explain how neurons produce raw, meaning-independent feels, but how biological processes implement meaningful integration—the kind of challenge science routinely addresses (how temperature arises from molecular kinetic energy, life from chemistry, meaning from computation).

Asking "why should meaningful integration feel like anything at all?" presupposes that phenomenology could be absent from the recursive process that constitutes it—a category error the framework rejects. The question is dissolved, not merely narrowed, and traditional pseudoproblems fall away with it: "Why does red feel like *this*?" (there is no meaning-independent "this"), "How can we compare your red to mine?" (experiences are intersubjectively structured), "Could zombies exist?" (consciousness is functional integration, not ghostly extra).

The question mistakes a constitutive relationship for a causal one—like asking why a system at 373 K at standard pressure also has the property of being at the boiling point of water. The two properties are not separately instantiated; they are the same fact at different levels of description. Likewise the phenomenal and the architectural are the same fact described from outside and from within. The explanatory gap depended on treating phenomenology as

ontologically separate from functional architecture; once that separation is rejected, the gap closes.

Objection 4: Cultural Variation Threatens Universality

The traditional framework makes cultural variation puzzling. If emotions are biological universals with the same raw phenomenology everywhere, why do cultures categorize them so differently? Standard responses fail empirically: same emotions with different labels cannot explain why concepts don't translate; same core emotions plus overlay cannot specify which emotions are core; display rules over universal feelings cannot 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 quite have this feeling because English lacks this concept. The biological core affect might be similar (pleasant low-arousal with affiliative motivation), but the specific feeling *is* the categorizing of that core affect using *amae* with its particular meanings in Japanese social life.

This is not "just semantic." The concept shapes phenomenology by determining what situations trigger the feeling (expecting benevolence), what action tendencies it includes (seeking caregiving), what moral valence it carries (positive dependence vs. problematic neediness), and what developmental arc it follows. English speakers with 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.

The preceding objections targeted the framework's dissolution of traditional puzzles. The following engage RSP's positive claims more directly.

Objection 5: "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 that presents its output as "raw feel" even though the experience is already conceptually structured. When you introspect "red," you access visual experience already interpreted as color and surface property, already carrying conceptual associations (warmth, danger), semantic knowledge, and contextual meaning. 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.

Consider an analogy. You "directly experience" understanding language, as if meaning flows immediately from sounds, yet psycholinguistics reveals extensive unconscious phonological, syntactic, and semantic processing. This analogy has independent computational support: Barenholtz's (2025) analysis of large language models argues that linguistic meaning requires no fixed mapping to inner referents — relational generative structure alone suffices — which parallels RSP's claim that the felt "rawness" of experience is the fluent operation of generative integration, not access to a pre-existing qualitative given. (The inference from LLM behavior to human language is by parsimony, and Barenholtz's essay is not peer-reviewed, so this is convergent illustration rather than evidence.)

The claim that introspection misreports its own structure is also supported by empirical work on introspective accuracy. One case bears directly on the character of current experience: Schwitzgebel (2011) documents how unreliable introspective judgments about one's own ongoing phenomenology can be. Three others concern adjacent targets — the causes of behavior (Nisbett & Wilson, 1977); the grounds of a choice, where subjects fail to notice a covertly switched outcome and readily justify a preference they never expressed (choice blindness, a mismatch-detection paradigm; Johansson, Hall, Sikström, & Olsson, 2005); and the reach of one's own understanding, overestimated until the explanation is demanded and the confidence collapses (the illusion of explanatory depth; Rozenblit & Keil, 2002). None of these three is about felt character, but together they shift the burden: once introspection is shown to be constructive wherever it can be checked against an independent standard, the claim that it stays uniquely transparent in the one domain where no such check is available, phenomenal character

itself, needs a positive argument the objector has not supplied. The findings fit a common picture: introspection generates a report on demand rather than reading off a standing inner record. That is what RSP would lead us to expect if the felt "rawness" of experience is the fluent operation of the self-model rather than a readout of a pre-existing given (see the comparison with Chater's introspection-as-illusion thesis, Part II, where RSP endorses the confabulating self-model as the constitutive process rather than an illusion masking a hidden interior).

Objection 6: "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 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 music without temporal patterns, meaning without relationships between symbols, digestion without a digestive system processing food? In each case the supposed "extra" property is not separate from the structure; it is the structure from a particular perspective.

Phenomenology is what recursive self-modeling is like from inside the system executing it. When a system models its own states, uses those models predictively, feeds predictions back through recursion in service of social navigation (modeling self-as-perceived-by-others), and makes this accessible to action-selection and planning, that system has phenomenology by virtue of this structure. The feeling is intrinsic to being the system executing these recursive loops.

The zombie argument collapses: The 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, then a being with our recursive architecture would have phenomenology by virtue of that architecture, and a being lacking phenomenology would necessarily have a different architecture. The scenario is incoherent — it asks for a running recursive self-model that somehow lacks first-person perspective, like asking for a computation that does not compute.

The conceivability argument (Chalmers, 1996, 2010) holds that zombie worlds are conceivable because phenomenal concepts have distinct primary and secondary intensions: the secondary

intension of "pain" picks out C-fibers (or RSP Level 3b) in the actual world, but the primary intension — phenomenal-feel-as-such — is conceivably distinct from any functional state. The architectural functionalist denies this two-dimensional structure for phenomenal concepts. To possess the concept "pain-as-it-feels," a thinker must have the architecture running and must register, via the recursive self-model, the architecture's own states as belonging to the thinker. That registration gives phenomenal concepts their first-personal grip; it is the same operation by which the architecture constitutes phenomenology. So phenomenal concepts have no primary intension independent of the architecture. What looks like a residual "phenomenal feel" intension floating free of any architectural state is itself the output of the recursive self-model representing its own states as intrinsically qualitative — the meta-problem in action (Chalmers, 2018). A zombie, by stipulation, lacks the architecture, so it also lacks the phenomenal concepts in terms of which the scenario is supposed to be coherent. Conceiving a zombie is not conceiving a system identical in architecture but different in phenomenology; it is misdescribing a system's architecture with one's own phenomenal concepts.

The transcendental reading of RSP (§III) dissolves the argument at the root. Suppose 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 conscious — so it is not a zombie. If we insist Level 3b can operate without a subject position, we posit something additional to the functional architecture — but then that addition becomes an empirical target rather than a metaphysical mystery. Either way the zombie argument loses its force: it fails because Level 3b entails the subject position, or it becomes a testable question about what beyond Level 3b might be needed.

RSP's approach has a close relative in Carruthers' (2000) higher-order thought theory. RSP extends Carruthers by specifying the *content* of higher-order representations (social-modeling predictions applied to self) and grounding them in a developmental-evolutionary account.

RSP also addresses what Chalmers (2018) calls the "meta-problem": why we intuit that consciousness is hard to explain. The recursive self-model represents its own processing as having intrinsic phenomenal character because first-person access to the self-model is structurally different from third-person description. The "hardness" intuition is itself a prediction of the architecture.

Why it feels like something specific: Different feelings (anger vs. joy vs. fear) arise from different meaningful integrations 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 aggression, attention focused on the offender, self-

schema as wronged party, social prediction of confrontation, and recursively modeling yourself being angry and others perceiving your anger. This pattern *constitutes* what anger feels like. There's no additional "anger-quale" floating independently; the phenomenology is the pattern experienced from within.

Objection 7: "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 separable from function. 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 system must (i) maintain valenced states whose targets are set by its own regulation of its internal condition, not by externally supplied reward at training time; (ii) sustain a self-model that persists and is updated by ongoing experience, not reinstantiated stateless per interaction; (iii) deploy that self-model predictively in real time so its predictions feed back into ongoing predictive operations, not consulted as a descriptive log; (iv) implement recursion such that the recursive operations modify the structure performing them — biologically, synaptic plasticity, but in principle any mechanism (online weight updates, test-time training, structural reconfiguration) making recursion self-modifying rather than fixed at deployment; and (v) make the recursive predictions accessible to action-selection that affects the system's own internal state, not only its external outputs. These criteria are agnostic about carbon vs. silicon but not about architecture, and each is in principle measurable — so the position stakes itself on empirical signatures testable as the relevant systems are built.

Application to contemporary AI. As of 2026, candidate systems (large language models, reasoning-tuned models, agentic systems with external memory, embodied vision-language-action models, world-model architectures) do not satisfy criteria (i)–(v) in the form required. For LLM-based systems, feedback signals shape the policy but do not constitute valenced states tracking the system's own condition, and self-representations are confined to an inference window or stored as external text. Embodied and world-model architectures have genuine sensorimotor closure — the strongest current candidates for criterion (v) — but endogenous valence (i) and self-modifying recursion (iv) remain unmet, since the homeostatic ground for

valence is absent and weights are typically fixed at deployment. The position therefore predicts that current AI systems are not conscious and specifies which structural changes would warrant revising that prediction: a falsifiable claim about architecture, not a substrate prejudice.

Recent discussions converge on these criteria from other directions. Chalmers (2023), assessing whether large language models could be conscious, enumerates functional obstacles (recurrent processing, a global workspace, unified agency, world- and self-models) that track criteria (ii)–(v), and shares the substrate-neutrality stated above; because he brackets the hard problem, however, his assessment remains a credence over a checklist, ending on a residual question, whether a system with all these properties might still fail to be conscious, that the constitutive-identity thesis closes rather than leaves open. Guitchounts (2026) offers broad easy-problem competence as a proxy for moral consideration where RSP offers an architectural criterion, and the two diverge, since such competence can run without the recursive self-model (Objection 12). Nait El Barj (2026), who expressly declines to argue about machine consciousness, nonetheless converges on the substance of criterion (i), the homeostatic-grounding condition, from phenomenology rather than architecture, locating the gap in a system that "has never been hungry," though she ties it to biological embodiment where RSP treats it as architecturally closable. (These are the present reading of each discussion, not claims their authors endorse.)

This also dispatches Searle's (1980) Chinese Room and Searle's (1992) biological naturalism. The Chinese Room stipulates a system lacking criteria (i)–(v) and is therefore not a counterexample — it is exactly the kind of system architectural functionalism predicts to be unconscious. Biological naturalism — that consciousness depends on causal powers specific to biological neurons — is the substrate-essentialist claim; but the causal-explanatory regularities accounting for conscious behavior are stated at the architectural level and are multiply realizable (Putnam, 1967; Fodor, 1974). The biological naturalist owes an account of which biological property does the constitutive work; "specific biochemistry" without specification is a placeholder, not an answer.

Systems lacking features (i)–(v) do not merely lack phenomenology — they lack the recursive self-modeling functions themselves. Under Paper III, Proposition E.4, recursive self-models satisfying specified conditions converge to a fixed-point self-attractor, which the framework requires for the strange loop to constitute a stable subject position.

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 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; a mammal with Level 2 schemas has felt embodiment, orientation, hunger, and satiation. These are not "mere information processing" but genuine, simpler forms of phenomenality. 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 transforms experience from something the system *undergoes* into something it *knows it undergoes*. 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 graduated emergence. Mirror self-recognition, typically 18–24 months in WEIRD samples (Amsterdam, 1972; Lewis & Brooks-Gunn, 1979), marks a rudimentary self-schema at Level 2's upper boundary; cross-cultural variation in this milestone (Broesch et al., 2011) suggests the task may measure task-validity differences across cultures rather than universal Level 2 architecture (see Rochat, 2003). Looking-time studies have reported implicit false-belief tracking from as early as 7 months (Kovács et al., 2010) and 15 months (Onishi & Baillargeon, 2005), but this paradigm has a contested replication record: Kulke et al. (2018) failed to replicate systematically, and Heyes (2014) argued the pattern admits a non-mentalistic teleological-action-prediction interpretation (a constructivist account that, as Apperly & Butterfill's (2009) two-systems framework formalizes, is broadly compatible with the distinction here). This produces an interpretive fork. If the looking-time effects require belief representation, they are early implicit Level 3b precursors emerging in infancy; if they reflect sophisticated Level 2 schema-mediated action anticipation (per Heyes, 2014), they are Level 2 reaching its upper limits. The framework is compatible with either reading, and current evidence does not settle which is correct. Explicit first-order false-belief tasks are reliably passed around 4–5 years (Wimmer & Perner, 1983; meta-analysis in Wellman et al., 2001); fully recursive second-order theory of mind consolidates later, most children passing standard second-order tasks around 6–7 years and after (Perner & Wimmer, 1985). The strange loop of mature self-conscious phenomenality therefore consolidates gradually over an extended window, not at a threshold.

The constitutive identity claim, refined. The claim that phenomenology is recursive self-modeling applies specifically to the self-aware, self-referential dimension distinctive of mature

human experience. It does not deny simpler phenomenality at Levels 1-2; what fish and toddlers 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 levels. This preserves what is right about the original sharp formulation (Level 3b is genuinely distinctive) while avoiding its false implication (that creatures without Level 3b have no phenomenality at all). The difference is one of degree and kind within phenomenality, not a boundary between phenomenality and its absence.

Objection 8: "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 RSP they are not. 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 has never *instantiated* it for this input. When she sees red, her brain performs a meaningful integration it has not performed before — interoceptive response to red light (Level 1), schema-organized spatial-chromatic experience (Level 2), 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.

This goes beyond the phenomenal concept strategy (Loar, 1990; Papineau, 2002) and the ability hypothesis (Lewis, 1988; Nemirow, 1990). The former says Mary acquires a new concept but no new fact; the latter, a new ability but no new propositional knowledge. Both concede something new happens but locate it outside the space of physical facts. RSP locates it inside: what Mary acquires is the running of a specific recursive integration over her core affect, body schema, attentional schema, and self-schema, conditioned on chromatic input she has not previously processed — itself a physical process, fully describable in Paper III's framework. Her prior knowledge described the process; her post-experience knowledge includes its actual running in her own architecture — an architectural state her prior knowledge described but did not

constitute. Where the rivals fail to specify *which* concept or ability is acquired, RSP specifies the architectural process, predicts that Mary's experience will be shaped by her conceptual repertoire (rich vs. impoverished color concepts differ phenomenologically), and grounds the answer in a constitutive identity rather than a residue.

Objection 9: "What About the Cogito? Doesn't It Establish a Cartesian Self?"

The objection: Descartes' (1641/1996) cogito argues that the activity of doubting or thinking establishes the existence of a thinking subject. The argument seems to require a Cartesian *res cogitans* — a thinking thing that exists prior to and independently of any physical process. RSP makes the self constitutively dependent on a physical recursive architecture. But the cogito appears to show that the self is foundationally given to itself before any architectural facts are settled. Doesn't this falsify the constitutive claim?

Reply: The cogito establishes less than is sometimes attributed to it. It establishes the impossibility of doubting one's own ongoing thinking activity while engaged in the doubting. It does *not* establish a substance — a *res cogitans* — over and above the activity. This distinction is consistent with a tradition from Hume's (1739/1978) bundle theory through contemporary cognitive science (Metzinger, 2003; Hofstadter, 2007).

On the RSP reading, the cogito's force is preserved but its metaphysical commitments are reduced. The "I" in "I think, therefore I am" is constituted by the recursive self-modeling activity the doubt manifests; the thinking is not evidence for an antecedent thinker but the activity through which the thinker exists at all. Descartes's mistake is to infer from the activity to a substance the activity supposedly reveals. Doubting establishes that the recursive self-model is currently running, not that a substance exists prior to and independent of that running.

This also explains why the cogito feels undeniable. The recursive self-model presents its own running as foundationally given because the running is the standpoint from which givenness is registered. Asking "is there an 'I' that thinks?" while the model is operative is asking the model whether it is operative; the answer cannot be no, because the question is posed by the very architecture it asks about. So the cogito does not establish a self prior to physical organization; it establishes that whenever the recursive organization runs, there is by that fact a perspectival standpoint. It remains a powerful observation about first-person epistemic immediacy, but does not entail the dualistic ontology often read into it.

Objection 10: "Isn't This Epiphenomenalism in Disguise?"

The objection: If consciousness is constituted by recursive self-modeling, and the recursive self-modeling is itself implemented by physical neural processes, then phenomenology

supervenes on the physical without doing any independent causal work. Whatever the neural processes cause, they cause; whatever the phenomenology adds is causally inert. RSP looks like epiphenomenalism with extra steps.

Reply: This objection is sharper than it first appears, and a casual constitutive-identity claim does not answer it. The full reply distinguishes two questions the objection runs together: (1) does phenomenology have causal work distinct from the underlying neural activity? and (2) is the architectural-organizational level of description causally indispensable, or could the same facts be stated entirely at the lower level?

To (1): there is no two-step causal economy in which neural activity does the work and phenomenology rides along. There is one process described at two levels. Mental causation is real because the architectural level is causally indispensable, not because phenomenology adds a second causal arrow atop the neural one.

To (2): the organizational pattern that constitutes phenomenology is multiply realizable — different neural implementations could instantiate the same recursive self-modeling architecture (the standard non-reductive physicalist move; Putnam, 1967; Fodor, 1974). The architectural level is indispensable because the regularities that explain behavior are stated there. Why a person in fear flees a predator is not exhaustively captured by "amygdala firings caused gastrocnemius contractions"; the architectural fact — a particular recursive integration of threat-detection, body schema, and predictive self-modeling — is part of the explanation, not a redundant description.

In its strongest form the objection is Kim's (1998, 2005) causal exclusion argument: if physical causation is complete, mental causation is either identical to physical causation or epiphenomenal. RSP accepts the first horn — mental causation *is* physical causation viewed at the architectural level — but this does not make phenomenology inert. It makes the architectural-level description irreducible to lower-level ones, just as thermodynamic explanations are irreducible to particle-level ones of the same system. Pressure in an ideal gas is multiply realizable in particle dynamics yet still does causal work: a balloon expands when heated both because molecules impart larger impulses on collision and because $PV = nRT$. Both are causal explanations; neither is epiphenomenal. So too a person flees both because specific neurons fire and because a recursive architecture integrating threat-detection with self-modeling runs.

The epiphenomenalist worry persists only if "real" causation must occur at the lowest physical level and all higher-level descriptions are mere shorthand. RSP rejects that: reduction would lose the multiply realizable patterns that do the explanatory work. The architectural level is

where the real explanatory regularities live, hence where phenomenology lives, hence why it is causally indispensable rather than epiphenomenal.

Objection 11: "RSP Cannot Explain the Distinctive Realness of Ordinary Experience"

The objection: Even granting RSP's machinery, ordinary waking experience has a distinctive felt *realness* that dreams, imagination, and virtual environments lack. A theory that reduces experience to predictive self-modeling cannot explain why genuine perception feels real in a way simulations do not — so it leaves out something essential.

Reply: The objection assumes "realness" is a fixed phenomenal property attaching to genuine perception and absent elsewhere. The evidence points the other way. Singleton and Fotopoulou (2026) argue that the sense of *presence* in virtual environments and the sense of *reality* in ordinary perception are conceptually, mechanistically, and phenomenologically continuous — both generated by the same hierarchical predictive architecture combining bottom-up interoceptive and exteroceptive signals with top-down modulation, sharing a frontolimbic substrate; presence in VR just is "the existing mechanisms of the sense of reality applied to a virtual setting." If so, realness is not a metaphysical surplus distinguishing genuine from simulated experience but a graded product of the predictive machinery, dialed up or down by multisensory consistency and the weighting of top-down prediction. RSP predicts exactly this: felt realness is Level 1 affective grounding integrated through the schema and recursive self-model levels, disruptable (depersonalization-derealization, treated in Paper II) or inducible (high-presence VR) without any change in the metaphysical status of what is experienced. The objection's intuition is preserved and explained; what is denied is that felt realness evidences a phenomenal property over and above the operation of the architecture. (Singleton and Fotopoulou's neuroimaging is preliminary, they bundle several distinct phenomena under one umbrella, and they are not aware of RSP; the convergence is independent support, not endorsement.)

Objection 12: "Unconscious Prediction Contradicts the Prediction–Consciousness Link"

The objection: Katlowitz et al. (2026) recorded from human hippocampus during propofol-maintained general anaesthesia using Neuropixels microelectrodes and showed that single units and local field potentials encode oddball discrimination, semantic features of natural speech, grammatical structure, and — at minimum — features of upcoming words sufficient for future-word decoding (the authors caution that this future-word decoding "does not necessarily imply active prediction beyond contextualization"). The effect grew over approximately ten minutes,

demonstrating representational plasticity in the unconscious state. The authors position this against "prominent theories of consciousness" that hold "sophisticated pattern recognition, semantic interpretation and predictive processing all require conscious access." The cited targets are Global Workspace Theory (Baars; Dehaene & Changeux), Integrated Information Theory (Tononi, Boly, Massimini, Koch), Recurrent Processing Theory (Lamme), and Dennett's account; RSP is not cited (the framework post-dates the paper's acceptance), but it makes commitments in the same neighborhood. If hippocampal predictive processing continues without consciousness, the link between prediction and phenomenology that prediction-based theories of consciousness rely on appears under pressure.

Reply: The objection trades on a missing distinction. RSP does not claim that *all* predictive processing constitutes consciousness. It claims that *Level 3 recursive social prediction* — the loop in which the system predicts how it is being predicted by others, and predicts that prediction, and so on — constitutes phenomenal consciousness. Level 1 prediction (interoceptive homeostasis) and Level 2 prediction (schema-driven sensory and semantic anticipation) operate without phenomenal consciousness in many ordinary contexts and continue under anaesthesia. The level distinction was never a notational convenience; it carries empirical commitments, and Katlowitz et al.'s data is consistent with one of the cleanest commitments the distinction makes.

What they report is hippocampal schema-level encoding of semantic and grammatical features, plus decodable representations of upcoming words during natural-speech listening. Whether this last constitutes "active prediction" in the strong sense is something the authors leave open; the weaker reading — robust contextual encoding sufficient for future-word decoding — suffices here. On either reading, hippocampal semantic encoding is a Level 2 schema phenomenon in RSP's taxonomy: the hippocampus participates in the schema network organizing auditory, semantic, and temporal features for downstream Level 3 recursion. RSP's cascade-control commitment requires inner loops (Level 1, 2) to stabilize before the outer loop (Level 3) is commissioned, so disrupting the outer loop need not disrupt the inner loops it depends on. On RSP's account (an interpretation Katlowitz et al. do not endorse or discuss), propofol-maintained anaesthesia disrupts the cortico-cortical recurrent loops implementing Level 3, while Level 1 (brainstem interoception, vitals) and Level 2 (sensory and hippocampal schema processing) persist on their own dynamics. The ten-minute representational plasticity is what an isolated Level 2 schema system should do: continue updating its internal model under exposure, absent recursive integration into a self-model that would render it conscious.

The framing therefore reverses. Katlowitz et al. is not a counterexample but a confirmation of one of RSP's cleanest positive predictions: under anaesthesia, Level 1 and Level 2 processing should survive in regions housing the relevant schemas even as Level 3 recursive integration is

selectively disrupted. (Whether the precise mechanism is loss of cortico-cortical recurrent dynamics is RSP's theoretical claim, not something Katlowitz et al. argue.) The hippocampus is one candidate region; primary sensory cortex is another (residual responses there under anaesthesia have long been reported, as the authors note). What was missing was a demonstration of *high-level* features — semantic and grammatical encoding, and future-word decoding consistent with online prediction — surviving in a region "anatomically and functionally distant from primary sensory cortices" (Katlowitz et al.). They supply that.

Two residual worries. First, that the Level 1/2/3 distinction is doing work just to absorb the result, so any prediction-based theory could relabel its commitments after the fact. But the distinction was not introduced for this purpose: the three-level architecture appears throughout the framework on independent developmental, clinical, evolutionary, and formal grounds (timescale separation τ_1 sub-second to early-second, $\tau_2 \approx 1-10s$, $\tau_3 \approx 10s-min$). The data lands on a partition already drawn. Second, that the reply renders RSP unfalsifiable — surviving prediction gets called Level 1/2, non-surviving Level 3. But the assignment is not free: Level 3 has architectural fingerprints (cortico-cortical recurrent loops involving prefrontal cortex, temporoparietal junction, and medial prefrontal regions tied to self- and other-modeling), and Katlowitz et al.'s hippocampal recordings do not target those structures. An experiment that targeted the Level 3 cortico-cortical loop and *still* found preserved recursive self-modeling under anaesthesia would be a genuine challenge. The current finding is not.

VII. Conclusion

What This Paper Has Accomplished

This paper has argued that the "hard problem" of consciousness dissolves once we abandon the raw qualia assumption. The Recursive Social Prediction model provides a three-level architecture in which phenomenal consciousness is constituted by — not produced by — physical processes organized in a specific recursive form. The constitutive-identity claim defended in §IV holds at the strong ontic level: phenomenology and the running of the architecture are one fact, accessed in two ways.

The dissolution works because it replaces a bad question ("Why should physical processes feel like something?") with a tractable one ("What organizational form constitutes phenomenal experience, and why is it structured this way?"). The RSP architecture answers the tractable question with specificity the hard problem never delivered.

The Objections

We addressed the major philosophical challenges to the dissolution. Jackson's Mary gains new meaningful integration when she sees red for the first time; she does not access a previously hidden quale. The zombie argument assumes what it needs to prove: that phenomenology is separable from the recursive self-modeling process. Descartes' Cogito, far from undermining the dissolution, exemplifies it: the certainty of "I think" is what a running strange loop registers from inside. The epiphenomenalism charge fails because the recursive self-model is causally indispensable at its level of description: remove it and the architectural-level explanations fail, even if the lower-level neural causal story remains formally complete. The functionalism charge is partially correct: RSP is a functionalism, but an architectural one, specifying a particular three-level recursive structure rather than generic computational sufficiency.

The Bat Reframed

Nagel's bat is not a philosophical mystery demonstrating the irreducibility of consciousness. It is an empirical challenge inviting comparative study of meaningful integration across species. The question is not "What is it like to be a bat?" but "What meaningful integrations does a bat's architecture perform, and how do they differ from ours?" That question is tractable. The evolutionary evidence (Part II) and mathematical formalization (Part III) show how to answer it.

What Comes Next

This paper is the philosophical foundation. Part II presents the evolutionary, developmental, and clinical evidence supporting the RSP architecture, along with seven falsifiable predictions. Part III provides the mathematical formalizations: four formal traditions (hierarchical predictive coding, variational free energy, intrinsically motivated reinforcement learning, and cascade control theory) unified by a single master functional, with conditional convergence theorems, stability analysis, and algorithmic specifications.

If the argument here is correct, consciousness research does not need a new bridge law between the physical and the phenomenal. It needs better descriptions of the organizational forms that constitute experience. That is a project for science, not metaphysics. The bat question should not haunt us. It should inspire us.

References

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

- Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4), 953-970.
- Ayer, A. J. (1940). *The foundations of empirical knowledge*. Macmillan.
- Barenholtz, E. (2025). LLMs show language does not describe reality. *IAI News*. <https://iai.tv/articles/llms-show-language-does-not-describe-reality-auid-3043>
- 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.
- Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in Experimental Social Psychology*, 41, 167-218.
- Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt.
- Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304), 1293-1295.
- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.
- Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5-6), 481-499.
- 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.
- Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge University Press.
- 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.

- Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*. <https://arxiv.org/abs/2303.07103>
- 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.
- 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.
- Dennett, D. C. (1988). Quining qualia. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 42-77). Oxford University Press.
- Descartes, R. (1641/1996). *Meditations on first philosophy* (J. Cottingham, Trans.). Cambridge University Press.
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178-190.
- Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese*, 28(2), 97-115.
- Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12), 11-39.
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127-138.
- Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Lawrence Erlbaum.
- Gallagher, S. (2005). *How the body shapes the mind*. Oxford University Press.
- Graziano, M. S. A. (2013). *Consciousness and the social brain*. Oxford University Press.
- Guitchounts, G. (2026). When the machines deserve our consideration. *Noema Magazine*. <https://www.noemamag.com/when-the-machines-deserve-our-consideration>
- Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, 17(5), 647-659.
- Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin & Review*, 22(2), 313-327.

- 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* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford University Press.
- Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127), 127-136.
- James, W. (1890). *The principles of psychology* (Vol. 1). Henry Holt.
- Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745), 116-119.
- Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the National Academy of Sciences*, 116(42), 20904-20909.
- Kant, I. (1781/1998). *Critique of pure reason* (P. Guyer & A. W. Wood, Trans.). Cambridge University Press.
- 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.
- Katlowitz, K. A., Cole, E. R., Mickiewicz, E. A., Shah, S., Franch, M., Adkinson, J. A., Belanger, J. L., Mathura, R. K., Meszéna, D., McGinley, M., Muñoz, W., Banks, G. P., Cash, S. S., Hsu, C.-W., Paulk, A. C., Provenza, N. R., Watrous, A. J., Williams, Z., Goldman, A. M., Krishnan, V., Maheshwari, A., Heilbronner, S. R., Kim, R., Rungratsameetaweemana, N., Hayden, B. Y., & Sheth, S. A. (2026). Plasticity and language in the anaesthetized human hippocampus. *Nature*. <https://doi.org/10.1038/s41586-026-10448-0>
- 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.
- Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, 330(6012), 1830-1834.
- Kripke, S. A. (1980). *Naming and necessity*. Harvard University Press.
- 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.

- Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(6), e1503.
- Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, 29(6), 888-900.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64(4), 354-361.
- Lewis, D. (1988). What experience teaches. *Proceedings of the Russellian Society*, 13, 29-57.
- Lewis, M., & Brooks-Gunn, J. (1979). *Social cognition and the acquisition of self*. Plenum Press.
- Loar, B. (1990). Phenomenal states. *Philosophical Perspectives*, 4, 81-108.
- Locke, J. (1689/1975). *An essay concerning human understanding* (P. H. Nidditch, Ed.). Oxford University Press.
- Medford, N., Sierra, M., Stringaris, A., Giampietro, V., Brammer, M. J., & David, A. S. (2016). Emotional experience and awareness of self: Functional MRI studies of depersonalization disorder. *Frontiers in Psychology*, 7, 432.
- Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity*. MIT Press.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.
- Nait El Barj, H. (2026). How AI will change us. *Noema Magazine*. <https://www.noemamag.com/how-ai-will-change-us>
- Nemirow, L. (1990). Physicalism and the cognitive role of acquaintance. In W. Lycan (Ed.), *Mind and cognition* (pp. 490-499). Blackwell.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, 308(5719), 255-258.
- Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14(1), 30-80.
- Papineau, D. (2002). *Thinking about consciousness*. Oxford University Press.

- Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a "theory of mind." *Philosophical Transactions of the Royal Society B*, 362(1480), 731-744.
- Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that...": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3), 437-471.
- Powell, L. E., Isler, K., & Barton, R. A. (2017). Re-evaluating the link between brain size and behavioural ecology in primates. *Proceedings of the Royal Society B*, 284(1865), 20171765.
- Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37-48). University of Pittsburgh Press.
- Rochat, P. (2003). Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition*, 12(4), 717-731.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521-562.
- Russell, B. (1912). *The problems of philosophy*. Williams & Norgate.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.
- 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.
- 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.
- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 42, 9-34.
- Schwitzgebel, E. (2011). *Perplexities of consciousness*. MIT Press.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424.
- Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

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

Sierra, M., & David, A. S. (2011). Depersonalization: A selective impairment of self-awareness. *Consciousness and Cognition*, 20(1), 99-108.

Singleton, O., & Fotopoulou, A. (2026). Presence is reality: Rethinking virtual and real-world consciousness. *Perspectives on Psychological Science*, 21(3), 269-291.
<https://doi.org/10.1177/17456916251414030>

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.

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.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1), 103-128.