Theories of Consciousness

Constraints, Commitments, and Convergence

Project: Return to Consciousness
Author: Bruno Tonetto
Authorship Note: Co-authored with AI as a disciplined thinking instrument—not a replacement for judgment. Prioritizes epistemic integrity and truth-seeking as a moral responsibility.
Finalized: March 2026

Abstract

This essay applies the constraint-based methodology of *Integration by Constraints* to major theories of consciousness — IIT, Global Workspace, predictive processing, Higher-Order Theories, Recurrent Processing, Attention Schema Theory, Orch OR, and others. For each theory, the essay separates structural discoveries about consciousness (constraints) from inherited ontological claims (commitments — usually physicalism). Five recurring structural features emerge across the theory landscape: integration, global accessibility, self-reference, anticipatory modeling, and non-trivial unity. The essay examines whether these features are ontologically neutral or create differential pressure between frameworks, maps where each theory's explanation terminates (its brute facts), and assesses the cost of reinterpreting constitutive identity claims under alternative ontologies. Theories are grouped into three categories: those identifying structural constraints (most neuroscientific theories), those denying the explanandum (illusionism and strong deflationary positions), and those relocating the primitive (panpsychism, Russellian monism, cosmopsychism). The analysis concludes that the apparent conflict between consciousness science and consciousness-first metaphysics is largely an artifact of ontological packaging rather than of the empirical findings themselves. The comparative assessment — that analytic idealism accommodates the convergent findings with fewer unexplained transitions — is explicitly framed as a theoretical-virtue comparison, the standard by which all ontological assessments proceed.

Keywords: theories of consciousness · constraint-based reasoning · integrated information theory · global workspace theory · predictive processing · illusionism · panpsychism · ontological neutrality · analytic idealism · philosophy of mind

I. The Question Behind the Theories

The science of consciousness has matured remarkably in the past three decades. What was once dismissed as philosophically intractable — or worse, scientifically disreputable — now sustains major research programs, dedicated journals, international conferences, and adversarial collaborations designed to adjudicate between competing frameworks. The landscape is rich: Integrated Information Theory, Global Neuronal Workspace Theory, Higher-Order Theories, Recurrent Processing Theory, predictive processing accounts, Attention Schema Theory, Orchestrated Objective Reduction, and several others each offer distinct accounts of what consciousness is and how it arises.

These theories are typically presented as competitors. IIT and GNW were pitted against each other in the Templeton Foundation's adversarial collaboration (2023). Illusionism positions itself against any theory that takes phenomenal consciousness at face value. Panpsychism challenges the shared assumption of most neuroscientific theories that consciousness emerges at some level of complexity.

But what exactly are these theories competing about?

This essay argues that the competition is less deep than it appears. Most theories of consciousness are doing something more specific — and more valuable — than settling ontology. They are identifying **structural constraints**: features of consciousness that any adequate account must explain. Integration, global access, self-modeling, recurrence, hierarchical prediction — these are structural findings about how consciousness is organized. They constrain theory. They do not, by themselves, determine what consciousness fundamentally is.

The apparent competition arises because most theories package their structural discoveries inside an ontological commitment — usually physicalism — that is not entailed by the discoveries themselves. When IIT says "consciousness is integrated information," the finding is that integration characterizes conscious systems. The additional claim — that information integration in physical substrates *generates* consciousness — is an ontological commitment that goes beyond what the structural finding establishes. When GNW says "consciousness is global broadcasting," the finding is that conscious content is globally accessible. The additional claim — that broadcasting in neural networks *produces* experience — is metaphysics, not mechanism.

This is not a criticism of these theories. It is a diagnostic clarification. As *The Generativity Question* argues, scientific theories are ontologically portable — their predictive and explanatory content does not depend on the ontological interpretation placed upon them. This essay applies that insight systematically to the major theories of consciousness.

Analytic idealism — the position, developed most systematically by Bernardo Kastrup, that reality is fundamentally mental and that individual minds are dissociated segments of a transpersonal consciousness — is the framework this project works within. It is argued for in *Return to Consciousness* and assumed here. The methodology is that of *Integration by Constraints*: separate what must be explained (constraints) from how it is interpreted (commitments). The result is a map of what consciousness science has found — and an argument that most of those findings are not only compatible with analytic idealism but are accommodated by it with fewer unexplained transitions than physicalism requires.

II. The Analytical Framework

Constraints vs. Commitments in Consciousness Science

Integration by Constraints distinguishes two levels of theoretical claim:

Constraints are features of the phenomenon that any adequate account must explain. They are discovered, not chosen. A constraint says: "Whatever consciousness is, it exhibits this feature." Some constraints sit comfortably within multiple frameworks — integration, for instance, is equally at home in physicalism and idealism. Others create differential pressure: they are accommodated more naturally by one framework than another, or resist accommodation by one entirely. Both kinds are genuine constraints — features of the territory, not artifacts of interpretation. What makes something a constraint is that it is a discovery about the structure of consciousness, not a claim about what consciousness ultimately is.

Commitments are claims about the ultimate nature of the phenomenon. They are adopted, not discovered. They position the theorist within a specific ontological framework. A commitment says: "Consciousness is fundamentally this kind of thing." The claim that "consciousness involves integration" is a constraint; the claim that "integration in physical substrates *produces* consciousness" is a commitment — it adds an ontological direction that the structural finding does not establish.

Most theories of consciousness contain both. The task is to separate them — not to diminish the theories, but to identify where structural discovery ends and ontological interpretation begins.

The Portability Test

The Generativity Question establishes that scientific theories are ontologically portable: their predictive content transfers intact across different metaphysical frameworks. The predictions of quantum mechanics are the same whether reality is fundamentally physical, mental, or neutral. The same applies to theories of consciousness: the structural findings — about integration, broadcasting, prediction, recurrence — describe features of consciousness that hold regardless of what consciousness ultimately is.

The test for each theory is: **Is this a discovery about the structure of consciousness, or a claim about what consciousness fundamentally is?** Structural discoveries are constraints. They may be ontologically neutral — compatible with both physicalism and idealism — or they may create asymmetric pressure, fitting more naturally within one framework than the other. In either case, they describe the territory. Ontological claims are commitments. They interpret the territory rather than map it — and they are what the theories disagree about most deeply while often presenting them as findings.

Three Categories

Applying this analysis across the landscape of consciousness theories reveals three categories:

1. **Theories that identify structural constraints.** Their core findings describe features of consciousness that are ontologically neutral. Most major neuroscientific theories fall here.
2. **Theories that challenge the explanandum.** They deny that phenomenal consciousness — the "what it is like" quality — exists as traditionally conceived. These are genuinely incompatible with any framework that takes experience as fundamental.
3. **Theories that relocate the primitive.** They modify what counts as fundamental in ways that converge with or approximate idealism. These are not competitors to idealism but neighbors.

The following sections examine each category in detail.

III. Theories That Identify Structural Constraints

Integrated Information Theory (IIT)

Core claim: Consciousness is identical with integrated information (Φ). A system is conscious to the degree that it is both differentiated (capable of many possible states) and integrated (cannot be decomposed into independent subsystems without information loss). The theory provides a mathematical formalism for quantifying consciousness.

The structural finding: Conscious systems exhibit high integration — their parts are informationally connected in ways that cannot be reduced to the sum of independent contributions. This is a genuine discovery about the *structure* of consciousness. It constrains any adequate theory: an account that cannot explain why conscious experience is unified rather than fragmented fails to cover the territory.

The ontological packaging: IIT identifies Φ with consciousness itself — not as a correlate or measure but as an identity. This is a strong ontological commitment: consciousness *is* integrated information, wherever it occurs. The commitment carries further implications: any system with nonzero Φ is conscious (panpsychism follows), and certain grid structures could in principle be more conscious than brains.

Portability assessment: The structural finding — that integration characterizes conscious systems — transfers intact to idealism. If consciousness is fundamental, we would expect it to exhibit integration as a native property. Mental life *is* integrative: thoughts cohere, perceptions bind, narratives unify. Under idealism, IIT's Φ becomes a measure of how consciousness organizes itself — an index of experiential integration, not a generator of it. The mathematical formalism is preserved; what changes is whether Φ is constitutive of consciousness or descriptive of its structure.

What does not transfer is the identity claim: that Φ *is* consciousness in a substrate-independent sense. Under idealism, Φ is the extrinsic appearance of how consciousness integrates — how integration *looks from outside* when measured in information-theoretic terms. The mathematical structure is real and informative; the ontological identification of structure with experience is the commitment that goes beyond the finding.

What IIT actually contributes as constraint: Consciousness involves integration. Any adequate account must explain why experience is unified rather than fragmented, and why disruptions of integration (as in split-brain cases, dissociative disorders, or anesthesia) alter the character of experience. This constraint is robust across methods, recurrent across contexts, resistant to eliminative explanation, and costly to exclude — meeting all four IBC criteria.

Global Neuronal Workspace Theory (GNW)

Core claim: Consciousness arises when information is broadcast globally across a "workspace" of interconnected cortical areas — primarily prefrontal and parietal networks. Unconscious processing is local and modular; conscious processing is what achieves global availability for report, reasoning, and flexible behavioral control.

The structural finding: Conscious content is globally accessible. There is a qualitative difference between information that remains locally processed and information that becomes available system-wide. This transition — from local to global — correlates with the transition from unconscious to conscious processing. The finding is supported by extensive experimental evidence: attentional blink paradigms, masking studies, and the neural signatures of reportable versus unreportable stimuli.

The ontological packaging: GNW is typically presented within a physicalist framework: global broadcasting in neural networks *produces* consciousness. Some formulations go further, suggesting that the hard problem dissolves once access consciousness is explained — that there is nothing more to consciousness than global availability for report and control.

Portability assessment: The structural finding — that consciousness involves global accessibility — transfers directly to idealism. Under analytic idealism, the brain is the extrinsic appearance of a dissociative process that localizes consciousness. Global workspace broadcasting would be the neural correlate — the *appearance from outside* — of how consciousness makes content available across its own structure. The workspace is not generating experience; it is the way experience's self-organization appears when observed through neuroimaging.

What does not transfer is the production claim: that broadcasting produces experience. Under idealism, broadcasting is what integration *looks like* in the brain — the extrinsic appearance of consciousness distributing content to itself. The mechanism is real; the ontological interpretation is the commitment.

The claim that access consciousness exhausts consciousness — that there is nothing more to explain once global availability is accounted for — is a stronger commitment that conflicts with any framework taking phenomenal consciousness seriously. But this is not a finding of GNW; it is a philosophical position adopted by some of its proponents (notably Dehaene). The empirical program stands without it.

What GNW actually contributes as constraint: Consciousness involves global availability. Content that is merely processed locally — without entering the global workspace — is not consciously experienced (or at minimum, not reportably so). Any adequate account must explain why consciousness has this broadcast structure rather than remaining locally encapsulated.

Recurrent Processing Theory (RPT)

Core claim: Consciousness correlates with recurrent (feedback) processing in sensory cortex, not with frontal broadcasting. Feedforward sweeps through the cortical hierarchy are unconscious; recurrent loops — where higher areas feed back to lower areas — generate phenomenal experience, potentially even without global access or reportability.

The structural finding: Recurrence — self-referential processing where outputs loop back to become inputs — is a feature of conscious processing. This finding cuts against GNW's emphasis on frontal broadcasting and suggests that consciousness may be more closely tied to posterior cortical dynamics than to prefrontal-parietal workspace activity. Evidence from change blindness, inattentional blindness, and the preserved phenomenology of certain lesion patients supports this.

The ontological packaging: RPT is typically neutral on deep ontology — Lamme himself focuses on the neural architecture rather than metaphysical questions. But insofar as recurrence is presented as the *mechanism that produces* phenomenal consciousness (rather than as its neural correlate), a physicalist commitment is implicit.

Portability assessment: Recurrence is straightforwardly compatible with idealism. If consciousness is fundamental, self-reference is not a surprising property — it is arguably constitutive. Awareness that is aware of itself, experience that loops back on itself, cognition that monitors its own processes — these are native properties of mentation. Under idealism, recurrent processing in the brain is the extrinsic appearance of consciousness's inherent self-referential character. The neural architecture reflects a feature of mind, not a mechanism that generates it.

What RPT actually contributes as constraint: Consciousness involves self-reference or recurrence. The transition from unconscious to conscious processing correlates with the transition from feedforward to recurrent dynamics. Any adequate account must explain why consciousness has this reflexive character.

Predictive Processing and Active Inference

Core claim: The brain is fundamentally a prediction machine — a hierarchical generative model that continuously generates predictions about sensory input and updates itself via prediction errors. Consciousness is what it is like to be such a model: perceptual experience is a "controlled hallucination" (Seth), and selfhood arises from interoceptive predictions about the body.

The structural finding: Conscious experience is structured by prediction. Perception is not passive registration but active inference — the brain constructs models and updates them when predictions fail. This framework unifies perception, action, attention, and selfhood under a single computational principle. The evidence is extensive: predictive coding accounts explain perceptual illusions, attentional modulation, placebo effects, and the phenomenology of psychedelic states (where prediction error increases and top-down models destabilize).

The ontological packaging: Predictive processing is often presented as a physicalist account: the brain's predictive computations produce conscious experience. Anil Seth's "beast machine" framework is more nuanced — Seth engages the hard problem directly and acknowledges it remains open — but the default framing is still that prediction produces experience rather than describes its structure.

Portability assessment: Prediction transfers to idealism exceptionally well. Mental life *is* predictive: anticipation, expectation, surprise, learning — these are not properties that need to be derived from neural computation; they are characteristic of mentation as such. Under idealism, the brain's predictive architecture is the extrinsic appearance of how consciousness structures its own experience — generating expectations, registering surprise, and updating its models of the world. The Bayesian mathematics is preserved; the ontological claim that computation produces experience is the commitment that goes beyond the finding.

The active inference framework (Friston) generalizes prediction to all self-organizing systems under a free energy principle. Under idealism, the free energy principle would describe how consciousness maintains its dissociative boundaries — the mathematical formalism capturing the dynamics of boundary maintenance that analytic idealism posits as fundamental. This is speculative but structurally coherent: the free energy principle as the extrinsic description of dissociative dynamics.

What predictive processing actually contributes as constraint: Consciousness is anticipatory and model-based. Experience is not passive reception but active construction. Any adequate account must explain why consciousness generates predictions, registers surprise, and updates its models — and why disruption of this predictive architecture (as in psychedelic states, psychosis, or sensory deprivation) alters the character of experience in characteristic ways.

Higher-Order Theories (HOT)

Core claim: A mental state is conscious when it is the object of a higher-order representation — a thought about the thought, or a perception of the perception. Without this higher-order monitoring, the first-order state is processed but not experienced.

The structural finding: Consciousness involves self-monitoring. There is something distinctive about mental states that are represented at a higher order — they have a quality of being "for me," of being experienced rather than merely processed. The theory draws support from cases where higher-order and first-order states dissociate: blindsight patients process visual information without experiencing it, suggesting that first-order processing without higher-order representation yields unconscious cognition.

The ontological packaging: HOT theories are typically physicalist: higher-order neural representations produce phenomenal consciousness. Some versions (notably Lau and Rosenthal's perceptual reality monitoring model) are more mechanistic, proposing specific neural circuits for higher-order monitoring.

Portability assessment: Self-monitoring is fully compatible with idealism. If consciousness is fundamental, awareness of awareness — meta-consciousness, reflexive knowing — is a native capacity, not something that must be constructed from non-conscious components. Under idealism, higher-order representation in the brain is the extrinsic appearance of consciousness monitoring itself. The neural architecture reflects a feature intrinsic to mind; it does not generate that feature from non-mental substrates.

HOT's implication that creatures without higher-order capacity are not conscious is the point where the theory's commitment matters most. Under idealism, this implication is reframed: creatures without higher-order neural architecture lack the dissociative structure that produces *individuated self-aware* experience, but they are not thereby excluded from consciousness altogether. The finding (self-monitoring characterizes human consciousness) is preserved; the commitment (no self-monitoring means no consciousness at all) is the part that requires ontological scrutiny.

What HOT actually contributes as constraint: Human consciousness involves a self-monitoring dimension — awareness of awareness. Any adequate account must explain why meta-cognition is closely tied to phenomenal consciousness and why its disruption (as in blindsight or certain dissociative states) alters experiential character.

Orchestrated Objective Reduction (Orch OR)

Core claim: Consciousness arises from quantum computations in microtubules within neurons. When quantum superpositions reach a critical threshold (related to quantum gravity), they undergo "objective reduction" — a non-computable process that is the physical basis of conscious moments.

The structural finding: Consciousness may involve non-computable processes — processes that cannot be captured by algorithmic computation. Penrose's mathematical arguments (from Gödel's incompleteness theorems) suggest that human mathematical understanding transcends what any Turing machine can achieve, implying that consciousness involves something beyond classical computation.

The ontological packaging: Orch OR locates consciousness in a specific physical substrate (quantum processes in microtubules) and a specific physical mechanism (objective reduction). This is a strong physicalist commitment — perhaps the most specific of any major theory.

Portability assessment: The non-computability finding, if correct, is compatible with idealism — and requires fewer auxiliary assumptions there than under physicalism. If consciousness is fundamental and non-computable, this is precisely what idealism would predict: experience is not the kind of thing that algorithms can generate, because it is ontologically prior to the computational structures that algorithms describe. Under idealism, Penrose's argument supports the claim that consciousness cannot be reduced to mechanism — which is a core idealist contention.

The microtubule hypothesis, by contrast, is a specific empirical claim about where to look for quantum coherence in biological systems. This is a scientific hypothesis testable on its own terms, independent of ontological commitments. Under idealism, quantum coherence in microtubules (if confirmed) would be the extrinsic appearance of fine-grained structure in how consciousness organizes itself at the neural level.

What Orch OR actually contributes as constraint: If its arguments are sound, consciousness may involve non-computable processes. Any adequate account must at least address the relationship between consciousness and computation — whether consciousness can in principle be fully captured by algorithmic processes. This remains contested but constitutes a serious constraint candidate.

Electromagnetic Field Theories (CEMI)

Core claim: Consciousness is the brain's electromagnetic field. Johnjoe McFadden's Conscious Electromagnetic Information (CEMI) field theory proposes that the EM field integrates information across neural populations and is the physical substrate of unified conscious experience.

The structural finding: Consciousness may be associated with field-level rather than neuron-level dynamics. The EM field is genuinely integrative — it binds information across spatially distributed neural populations in a way that individual synaptic connections do not. This could address the binding problem: how distributed neural processing gives rise to unified experience.

The ontological packaging: CEMI identifies consciousness with a physical field — the EM field is the substrate, not merely a correlate.

Portability assessment: The finding that consciousness correlates with field-level dynamics rather than discrete neural firings transfers to idealism. Under idealism, the EM field would be the extrinsic appearance of consciousness's integrative character — how experiential unity looks when measured electromagnetically. The identification of consciousness with the field is the commitment; the association between field dynamics and experiential unity is the constraint.

What CEMI actually contributes as constraint: Conscious experience may be unified at a field level rather than at the level of individual neural elements. Any adequate account must address why consciousness is unified and whether that unity corresponds to integrative physical processes at the field level.

IV. The Convergence Pattern

The theories examined in Section III differ substantially in their proposed mechanisms, preferred neural substrates, and empirical predictions. IIT emphasizes posterior cortex and integrated information; GNW emphasizes prefrontal-parietal broadcasting; RPT emphasizes sensory recurrence; predictive processing emphasizes hierarchical generative models; HOT emphasizes meta-representational circuits; Orch OR emphasizes quantum microtubule dynamics; CEMI emphasizes electromagnetic fields.

Yet when their structural findings are extracted from their ontological packaging, a pattern of recurrence emerges. Conscious experience, as these theories collectively describe it, involves:

1. **Integration.** Experience is unified rather than fragmented. Disruptions of integration alter or dissolve consciousness. (IIT, CEMI)
2. **Global accessibility.** Conscious content is available system-wide for flexible use — in reasoning, reporting, and behavioral control. (GNW)
3. **Self-reference.** Consciousness involves recurrence — processing that loops back on itself, monitoring its own states. (RPT, HOT)
4. **Anticipatory modeling.** Experience is not passive registration but active construction — prediction, inference, and updating. (Predictive processing)
5. **Non-trivial unity.** The binding of disparate elements into a single experiential field requires explanation — it is not an automatic consequence of having many processes running simultaneously. (IIT, CEMI, GNW)

These five features are not the discoveries of any single theory. They are structural findings that recur across competing theories — some independently discovered by multiple frameworks, others contributed distinctively by one but confirmed as relevant by the rest. Together they constitute the constraints that survive disagreement about mechanism and substrate. Any adequate account of consciousness must explain all five. These particular constraints are ontologically neutral — both physicalism and idealism can accommodate them. Section X examines a different class of empirical findings (psychedelics, anesthesia) that also function as constraints but create differential pressure between frameworks.

The Pattern Is Significant

This recurrence is evidentially significant, though its weight should not be overstated. The theories examined here share institutional context, methods, and training — they are not as independent as the cross-traditional convergence documented in *One Structure*. Still, when IIT and GNW agree that consciousness involves integration despite disagreeing about where integration matters most, the agreement is not trivially explained by shared method — it reflects the phenomenon constraining theory.

The convergence also has a specific character that matters for the ontological question. All five features — integration, global accessibility, self-reference, anticipatory modeling, and non-trivial unity — are **native properties of mind**. More precisely: they are native to *dissociated* mind — to bounded, finite consciousness operating under constraint. Under analytic idealism, individual minds are dissociated segments of universal consciousness, and it is these segments — not the undissociated whole — that anticipate, predict, integrate, and self-monitor. The features consciousness science has discovered characterize *what it is like to be a bounded experiencer*, which is exactly what the theories study. They are what mentation *does*:

- Mental life integrates: thoughts cohere, perceptions bind, narratives unify

- Mental content is globally available: a sudden insight reconfigures your entire understanding, not just the module that processed the data
- Mental processes are self-referential: you can think about thinking, attend to attention, be aware of awareness
- Mental life is anticipatory: expectation, surprise, planning, and learning are constitutive of experience
- Mental experience is unified: the phenomenal field is one — seeing, hearing, feeling, and thinking are not separate streams that happen to co-occur but aspects of a single experiential moment

Under physicalism, each of these features is an explanatory target — something that must be derived from non-mental processes through some combination of complexity, organization, and emergence. Each derivation requires crossing from the non-experiential to the experiential — the hard problem in miniature. Under analytic idealism, they are what consciousness looks like when it operates. No category crossing is required because the features are characteristic of the primitive itself.

This is a comparison of theoretical virtues — parsimony, explanatory coherence, the number of unexplained transitions — not a demonstration. But all ontological comparisons operate at this level. Physicalism's competing claim — that emergence from non-experiential processes "handles" consciousness — is equally a theoretical-virtue judgment, not an empirical proof. The question is which framework accommodates these findings with fewer structural costs, and the answer favors the framework within which those findings describe native features rather than emergent anomalies.

V. Theories That Challenge the Explanandum

Not all theories of consciousness aim to explain consciousness. Some aim to dissolve it — to show that the thing we thought needed explaining does not exist as conceived, or that our intuitions about it are systematically mistaken. These theories occupy a fundamentally different position from those examined in Section III. They are not identifying structural constraints on consciousness; they are denying that the apparent constraints are genuine.

Illusionism

Core claim: Phenomenal consciousness — the "what it is like" quality, qualia, the intrinsic character of experience — is an illusion. There are no qualia in the traditional philosophical sense. What exists is a set of functional and representational properties that the brain models as having intrinsic qualitative character. The sense that experience has an irreducible subjective quality is itself a product of how the brain represents its own states — not a feature of reality.

Keith Frankish distinguishes his position carefully: there is a real phenomenon to explain — why we *believe* there are qualia — but the qualia themselves are illusory. Daniel Dennett's earlier version ("quining qualia") argued that our intuitions about phenomenal consciousness are systematically confused and that the hard problem vanishes once the confusion is dissolved.

The constraint-based assessment: Illusionism does not identify a structural constraint on consciousness. It denies that the central constraint exists. Where every other theory examined in this essay takes the existence of phenomenal experience as a datum to be explained, illusionism treats it as a systematic error to be diagnosed.

This places illusionism in a unique position relative to the IBC framework. Recall the four criteria for constraint-candidacy: robustness across methods, recurrence across contexts, resistance to eliminative explanation, and cost of exclusion. The existence of phenomenal experience — that there is something it is like to see red, feel pain, or hear a melody — passes all four tests. It is reported across every method of investigation, recurs in every cultural and historical context, has resisted every eliminative attempt (since the denial itself is an experience), and its exclusion renders any account of consciousness visibly inadequate.

Illusionism's response is that the resistance to elimination is itself part of the illusion — the brain's self-model is designed to represent its own states as having irreducible qualitative character, so of course introspection reports irreducibility. The appearance of phenomenal consciousness is explained; the reality of it is denied.

This is a coherent philosophical position. It is also the position most directly incompatible with analytic idealism. If consciousness is the ontological primitive — if reality is fundamentally experiential — then denying the existence of phenomenal experience is not an interpretive disagreement but a foundational contradiction. Idealism and illusionism cannot both be correct; they disagree about whether the explanandum exists.

Why the incompatibility is genuine: Unlike the theories in Section III — whose structural findings describe features of consciousness that both physicalism and idealism can accommodate — illusionism's core claim does not describe a feature of consciousness at all. It denies the central one. You cannot hold that reality is fundamentally experiential and that experience is an illusion. The contradiction is not a matter of emphasis or interpretation; it is logical.

This does not settle which position is correct. It locates where the genuine disagreement lies. Illusionism and idealism share nothing at the level of the explanandum. The question between them is prior to all the mechanistic questions that occupy the other theories: *Is there something it is like to be conscious, or not?*

Attention Schema Theory (AST)

Core claim: The brain constructs a simplified internal model of its own attention processes — the "attention schema." Consciousness is this internal narrative about what attention is doing. We believe we have subjective experience because the model says we do, but the model is a simplification — like the body schema is a simplified model of the body.

Michael Graziano's theory exists on a spectrum. In its moderate form, AST identifies a genuine structural feature: the brain models its own attention. In its strong form, AST is deflationary — consciousness as traditionally conceived does not exist; what exists is a self-model that represents itself as being conscious.

The constraint-based assessment: The moderate form identifies a constraint: consciousness involves self-modeling, and attention modeling in particular is a feature of how conscious systems operate. This is compatible with idealism for the same reason as HOT — under idealism, self-modeling is a native property of mind.

The strong form — where the self-model is all there is and phenomenal consciousness is a representational fiction — is genuinely incompatible with idealism, for the same reason illusionism is. If consciousness is fundamental, it cannot be a fiction generated by a model.

The spectrum matters: AST's incompatibility with idealism depends on how strongly the deflationary claim is pressed. The structural finding (brains model their own attention) is a constraint. The denial of phenomenal consciousness is a commitment — and a commitment that contradicts any framework taking experience as ontologically basic.

What Deflationary Theories Reveal

The deflationary theories — illusionism and strong AST — serve an important diagnostic function even for theorists who reject their conclusions. They make explicit what is at stake in the foundational question: *Is phenomenal experience real?*

Every other theory examined in this essay — IIT, GNW, RPT, predictive processing, moderate HOT, Orch OR, CEMI — takes the existence of phenomenal experience as given and seeks to explain it. The deflationary theories refuse this starting point. In doing so, they clarify that the existence of the explanandum is itself a substantive commitment — one that idealism shares with most of consciousness science but that is not universally held.

The IBC framework has already addressed this. As established in *Integration by Constraints*, the existence of first-person experience passes all four constraint criteria, including resistance to eliminative explanation: the denial is self-refuting because the denial itself is an experience. Illusionism's reply — that the appearance of irreducibility is part of the illusion — is ingenious but encounters a structural problem: it requires a non-experiential system to generate the *experience* of being experiential, which reintroduces the hard problem at the meta-level. Explaining why the brain represents itself as having phenomenal states requires explaining what the representing *itself feels like* — and if it feels like nothing, the explanation explains something that by hypothesis does not exist.

This structural problem does not refute illusionism decisively. But it establishes that illusionism does not dissolve the hard problem — it relocates it. And it identifies the deflationary theories as the genuinely incompatible positions in the landscape, in contrast to the much larger set of theories whose findings are ontologically neutral.

VI. Theories That Relocate the Primitive

A third category of theories does not fit neatly into either the constraint-identifying or the explanandum-denying camps. These theories modify what counts as fundamental in ways that bring them into the neighborhood of idealism — sometimes close enough that the boundary between positions becomes unclear.

Panpsychism

Core claim: Consciousness is a fundamental and ubiquitous feature of reality. Rather than emerging at some level of complexity, experiential properties are present at all levels — including fundamental physics.

Variants:

- *Micropsychism* (Strawson, Goff): Fundamental physical entities have micro-experiences; macro-consciousness emerges from combinations of these.
- *Cosmopsychism* (Goff, Shani): The universe as a whole is the fundamental conscious subject; individual minds are derivative. The "decomposition problem" — how cosmic consciousness fragments into individual minds — replaces the combination problem.

- *Panprotopsychism* (Chalmers): Fundamental entities have proto-conscious properties that are not yet experiential but ground consciousness when appropriately organized.

Relationship to idealism: The relationship depends on the variant. Micropsychism retains the physicalist structure (particles are fundamental; consciousness is a property of particles) while modifying the content (particles are experiential). It faces the **combination problem** — how micro-experiences combine into unified macro-experience — which is structurally analogous to the hard problem it was designed to solve.

Cosmopsychism is structurally near-identical to analytic idealism. Both posit a fundamental consciousness of which individual minds are derivative aspects. The "decomposition problem" of cosmopsychism *is* idealism's "granularity problem" — why universal consciousness fragments into *these* specific individual minds. The frameworks converge on the same explanatory structure and face the same outstanding debt. The difference is primarily one of lineage and emphasis: cosmopsychism developed from analytic philosophy of mind; analytic idealism developed from continental and Schopenhauerian traditions via Kastrup's reformulation.

Panprotopsychism occupies a middle ground — its proto-experiential properties are neither fully experiential (which would make it panpsychism proper) nor fully non-experiential (which would make it physicalism). Whether proto-experience is closer to idealism or physicalism depends on how it is characterized — and the characterization remains underdeveloped, which is the position's principal weakness.

What panpsychism contributes as constraint: The intuition driving panpsychism — that deriving experience from non-experience is categorically problematic — is a constraint on any adequate account. As *First-Principles Assessment* argues, the emergence of consciousness from non-conscious processes constitutes a foundational inversion that physicalism has not resolved. Panpsychism registers this constraint and proposes that experience must be present at the fundamental level. Idealism agrees but goes further: not merely experiential *properties* attached to physical entities, but experiential *reality* as the ontological ground.

Russellian Monism

Core claim: Physics describes the world's relational and structural properties — causal roles, mathematical relations — but says nothing about the *intrinsic nature* of what occupies those roles. The intrinsic nature might be experiential or proto-experiential. On this view, consciousness is not emergent from something non-experiential; it is the intrinsic character of what physics describes extrinsically.

Relationship to idealism: *First-Principles Assessment* analyzes this in detail. Russellian monism dissolves the category-crossing problem that drives the hard problem — if the physical *is* experiential at its intrinsic level, there is no non-experiential-to-experiential transition to explain. But if this is correct, then "physicalism" in its standard sense — the view that reality is fundamentally non-experiential — is false. What succeeds is a view in which experience is fundamental, which is structurally closer to idealism than to the physicalism the label preserves.

The key question for Russellian monism is the combination problem: how do intrinsic experiential properties of fundamental entities combine into the unified macro-experience of a human mind? This problem is acute for micropsychist variants and less pressing for cosmopsychist variants — which brings us back to the convergence between cosmopsychism and idealism.

What Russellian monism contributes as constraint: Physics describes structure, not intrinsic nature. The relational character of physical description leaves the intrinsic nature of reality as an open question — one that physics itself cannot settle. This is a genuine constraint: any adequate account must explain the relationship between the structural properties physics describes and the intrinsic nature of what exhibits those properties. The constraint is compatible with idealism (the intrinsic nature is experiential), Russellian monism (the intrinsic nature is experiential or proto-experiential), and physicalism only if physicalism can provide a non-experiential account of intrinsic nature — which it has not done.

Neutral Monism

Core claim: Reality is fundamentally neither mental nor physical but some neutral "stuff" that can appear as either depending on how it is organized or viewed. Mind and matter are two aspects of the same underlying reality.

Relationship to idealism: Neutral monism avoids privileging either mind or matter. Its principal challenge is characterizing the "neutral" substrate — what is it, if neither mental nor physical? Depending on how this is resolved, neutral monism can collapse into physicalism (if the neutral base is characterized in non-experiential terms), idealism (if characterized in experiential terms), or remain genuinely distinct (if a third characterization is viable). The position's coherence depends on whether "neither mental nor physical" is a stable category or an unstable placeholder that must eventually resolve in one direction.

What neutral monism contributes as constraint: The duality of mental and physical descriptions may reflect perspectival difference rather than ontological division. Any adequate account must address why reality admits of both experiential and structural description — whether this dual character is fundamental (dual-aspect), derivative of something deeper (neutral monism), or the result of one being the appearance of the other (idealism and physicalism, from opposite directions).

VII. The Brute-Fact Map

Where Explanation Stops establishes that every framework must terminate somewhere — at brute facts that are accepted without further grounding. The question is not whether a framework has brute facts but where it places them and what that placement costs.

Applying this analysis to the theories of consciousness reveals a pattern:

Theory	Where It Stops	Type of Brute Fact
IIT	Integration produces consciousness	Why does Φ feel like anything?
GNW	Broadcasting produces consciousness	Why does global access generate experience?
RPT	Recurrence produces consciousness	Why does feedback processing feel like anything?
Predictive processing	Prediction produces consciousness	Why does being a predictive model feel like anything?
HOT	Higher-order representation produces consciousness	Why does meta-representation generate phenomenality?

Theory	Where It Stops	Type of Brute Fact
Orch OR	Quantum reduction produces consciousness	Why does objective reduction feel like anything?
CEMI	EM field dynamics produce consciousness	Why does a field feel like anything?
Illusionism	There is nothing to explain	Why do we experience being experiencers?
Panpsychism	Experience is fundamental	Why this combination? (combination problem)
Russellian monism	Intrinsic natures are experiential	Why this combination? (same problem)
Cosmopsychism	Cosmic consciousness is fundamental	Why this decomposition? (granularity problem)
Analytic idealism	Mind and dissociation are fundamental	Why this decomposition? (granularity problem)

Two patterns are visible.

Pattern 1: Every theory that takes consciousness as something to be *produced* by non-conscious processes faces a version of the hard problem at its brute-fact level. IIT, GNW, RPT, predictive processing, HOT, Orch OR, and CEMI all terminate at the same structural gap: *why does this process feel like anything?* The specific processes differ — integration, broadcasting, recurrence, prediction, meta-representation, quantum reduction, field dynamics — but the gap is the same. No amount of mechanistic elaboration closes it, because the gap is between structure and experience, not between simpler and more complex structure.

Pattern 2: Theories that take experience as fundamental face a different problem — the combination or granularity problem — but they do not face the hard problem. Panpsychism, cosmopsychism, Russellian monism, and analytic idealism all accept experience as primitive and ask how it organizes into the specific configurations we observe. This is a genuine debt, but it is a different *kind* of debt: a question about the structure of something whose existence is given, rather than a question about how something categorically new arises from something that lacks it entirely.

As *Where Explanation Stops* notes, physicalism starts from parts and must explain unity (the binding problem); idealism starts from unity and must explain parts (the granularity problem). The brute-fact map of consciousness theories reveals that this structural mirror operates across the entire landscape: theories that begin with non-experiential processes must explain the emergence of experience; theories that begin with experience must explain its specific configuration.

VIII. Idealism as Interpretive Home

The analysis to this point has been diagnostic — separating constraints from commitments, identifying convergence, and mapping brute facts. This section draws the implication.

What idealism does with the findings

Analytic idealism does not contest the empirical findings of any theory examined in this essay. It reinterprets their ontological significance:

Integration (IIT): Under idealism, integration is intrinsic to consciousness. Mind naturally unifies — thoughts cohere, perceptions bind, experience is whole. IIT's Φ measures the degree of this native integration as it appears in the brain's information-processing architecture. The mathematical formalism is preserved; the ontological direction is reversed.

Global accessibility (GNW): Under idealism, the global workspace is the neural appearance of how consciousness distributes content across its own field. Broadcasting is not generating experience; it is the extrinsic correlate of experience organizing itself. The clinical and experimental findings — attentional blink, masking, reportability thresholds — describe how dissociative boundaries structure what enters and exits conscious access.

Recurrence (RPT): Under idealism, self-reference is a fundamental property of consciousness. Awareness aware of itself — reflexive knowing — is what contemplative traditions have described for millennia (see *Reflexive Awareness*). Recurrent processing in sensory cortex is the neural appearance of this inherent self-referentiality.

Anticipatory modeling (predictive processing): Under idealism, prediction is native to mentation. Minds anticipate, expect, and update — not because they compute predictions but because anticipation is what mental activity *does*. The hierarchical predictive architecture of the brain reflects the hierarchical structure of how consciousness models its own experience.

Self-monitoring (HOT): Under idealism, meta-consciousness — awareness of awareness — is a developed capacity within consciousness, not something generated by higher-order neural circuits. The neural correlates of higher-order representation are the extrinsic appearance of consciousness's capacity to know itself.

Non-computability (Orch OR): Under idealism, the non-computable character of consciousness is expected — experience is not algorithmic because it is ontologically prior to computation. Penrose's arguments support idealism's contention that consciousness cannot be reduced to mechanism.

Field-level unity (CEMI): Under idealism, the electromagnetic field's integrative properties reflect the unity of consciousness as it appears in physical measurement. Experiential unity is fundamental; field-level integration is its extrinsic correlate.

What idealism does not do

Idealism does not:

- **Replace the science.** Every mechanism, every neural correlate, every experimental finding is preserved. What changes is the ontological interpretation, not the data.
- **Claim the theories were secretly idealist all along.** The theories were developed within physicalist frameworks, and their empirical programs are shaped by physicalist assumptions. Idealism reinterprets; it does not claim credit.
- **Dissolve the remaining questions.** Idealism faces its own debts — the granularity problem, the underdeveloped account of why *these* dissociative boundaries produce *these* experiential domains, and the need for research programs that generate testable theories within the space idealism opens. These are real and acknowledged.
- **Render the theories redundant.** The structural findings of consciousness science are valuable regardless of ontology. Understanding how integration, broadcasting, recur-

> rence, and prediction operate in the brain is genuine progress. Idealism does not make this work unnecessary; it provides a different interpretive frame for understanding what the work has found.

What the reinterpretation costs

Honesty requires naming what idealism's reinterpretation loses, not only what it preserves. Some theories — IIT most explicitly — make **constitutive identity claims**: Φ does not merely correlate with consciousness; Φ *is* consciousness. This identity gives IIT its distinctive explanatory force: it tells you that any system with high Φ is conscious, and it explains why disruptions of integration alter experience. When the essay reinterprets Φ as a measure of how consciousness organizes itself rather than as consciousness itself, the identity claim is stripped and IIT becomes a correlation — useful, but no longer the theory its proponents advanced.

This is a real cost. But it must be weighed against what the identity claim actually delivers. IIT's identification of Φ with consciousness is powerful *if* it can be grounded — if someone can explain why this particular mathematical structure is accompanied by experience rather than merely exhibiting it. That grounding has not been provided. The identity is asserted as an axiom, not derived from anything more fundamental. The same applies to every production-model identity: GNW's claim that broadcasting *is* conscious access, HOT's claim that higher-order representation *generates* phenomenality, predictive processing's claim that being a generative model *constitutes* experience. In each case, the identity is the theory's most ambitious claim and its least grounded one — because it is precisely where the hard problem lives.

What idealism's reinterpretation loses, then, is the *promissory force* of these identities — the prospect that mechanism-level description will eventually close the gap between structure and experience. What it does not lose is *demonstrated* explanatory content, because the gap has not been closed. The cost is real but it is the cost of an undelivered promise, not of an achieved explanation. Whether that promise will eventually be redeemed is an open question — but it cannot be counted as a current asset in a comparative assessment.

The structural-cost asymmetry

There is, however, an asymmetry in structural costs. Under physicalism, each structural feature of consciousness is an explanatory target — something that must be *derived* from non-conscious processes. Integration must emerge from non-integrated parts. Self-reference must arise from non-self-referential components. Prediction must be constructed from non-anticipatory substrates. Each derivation requires an unexplained transition from the non-experiential to the experiential.

Under idealism, these features are *native*. They do not need to be derived because they are characteristic of what consciousness is. The explanatory burden shifts from "How does this feature arise from non-mental processes?" to "How does this feature manifest in the brain's physical architecture?" — which is a question about extrinsic appearance, not about ontological generation.

This is a theoretical-virtue comparison: which framework accommodates the findings with fewer unexplained transitions? As *Asymmetric Methodological Restraint* argues, demanding empirical falsification for ontological comparisons — while accepting theoretical-virtue arguments when they favor physicalism — is itself an asymmetric standard. Physicalism's claim that emergence handles consciousness is no less a theoretical-virtue judgment than idealism's claim that consciousness features are native to mind. The difference is that one is treated as the default and the other as requiring special justification.

IX. The Production Assumption

A thread runs through Sections III–VIII that deserves explicit statement. Nearly every neuroscientific theory of consciousness contains a step that is metaphysical rather than empirical: the inference from *correlation* to *production*.

The Emergence of Physicalism identifies this pattern: "The correlation between brain states and mental states is an empirical finding; the production claim is a metaphysical addition that goes beyond the evidence." *Anomalous Phenomena and Consciousness* develops the point: lesion-deficit correlations establish dependence and modularity but not ontological production — a damaged filter and a damaged generator produce the same pattern of deficit.

The same analysis applies to each theory examined here:

- **IIT finds** that high integration correlates with rich consciousness. **It infers** that integration produces consciousness.
- **GNW finds** that global broadcasting correlates with conscious access. **It infers** that broadcasting produces consciousness.
- **RPT finds** that recurrent processing correlates with phenomenal experience. **It infers** that recurrence produces consciousness.
- **Predictive processing finds** that predictive dynamics correlate with experiential structure. **It infers** that prediction produces consciousness.
- **HOT finds** that higher-order representation correlates with phenomenal awareness. **It infers** that meta-representation produces consciousness.

In each case, the finding is empirical and robust. The inference is metaphysical and underdetermined. The correlation is a constraint; the production claim is a commitment.

This does not mean the production claim is false. It means it is not established by the findings that appear to support it. The same findings are equally compatible with the constraint model: the brain does not produce consciousness but constrains, filters, and structures it. Damage to the brain disrupts the constraint structure, producing specific experiential deficits — exactly as the production model predicts, but for different ontological reasons.

The production assumption is so deeply embedded in consciousness science that it is rarely stated as an assumption. It is treated as the default interpretation — the obvious reading of the data. But as this project has argued throughout (see *Myth of Metaphysical Neutrality* and *The Emergence of Physicalism*), the "obvious" reading is the product of a specific historical and institutional trajectory, not of the data themselves.

Making the production assumption visible does not refute it. It allows it to be evaluated as what it is: a metaphysical commitment, not an empirical finding. And it opens the space for the constraint model — where the brain is the extrinsic appearance of how consciousness structures itself — to be evaluated on equal terms.

X. The Anesthesia and Psychedelic Tests

Two domains of evidence provide particularly revealing tests of the framework developed here: anesthesia and psychedelics. Both involve pharmacological alteration of brain states with measurable changes in consciousness — and both generate patterns that the production model must accommodate with auxiliary hypotheses while the constraint model predicts naturally.

Psychedelics: More Experience from Less Activity

As *Anomalous Phenomena and Consciousness* documents in detail, psychedelic states are characterized by profoundly expanded experience — subjects report some of the most meaningful experiences of their lives — accompanied by *decreased* activity in key hub networks, notably the default mode network. The overall neural picture is complex (involving redistribution, altered connectivity, and increased entropy in some measures), but the core observation persists: the brain states most associated with reported expansion involve disintegration of the organized network architecture that ordinarily supports cognition.

If these findings are read through a production lens — where organized neural architecture *generates* experience — the pattern requires explanation, since disrupting that architecture should more often degrade experience than enhance it. The mechanism-level responses are available: IIT's formalism can describe entropy increases as reorganization of integrated information; GNW can point to altered broadcasting patterns; Carhart-Harris's REBUS framework within predictive processing specifically predicts that loosening hierarchical control can liberate content. These are legitimate neuroscientific accounts, ontologically portable in their own right. The production landscape is not monolithic on this point. But the core observation — that substantial network disintegration accompanies reported expansion — sits more easily within a framework that does not require organized architecture to *generate* experience in the first place.

Under the constraint model — the brain filters and structures consciousness rather than producing it — the general direction of the pattern is expected. Weakening the constraint structure permits access to content that is always present but normally filtered. The model does not predict the specific pharmacological details of which disruptions expand versus degrade experience — that requires the mechanism-level neuroscience both frameworks share. But its default expectation (weakened constraints, expanded access) aligns with the observed direction without requiring the kind of auxiliary accommodation that simple production models need.

Anesthesia: Consciousness Under Suppression

As *Conscious Under Anesthesia* documents, isolated forearm technique (IFT) studies show responsive awareness in up to 37% of patients during general anesthesia. Ketamine produces vivid phenomenology while satisfying clinical criteria for "anesthesia." The equation "anesthesia = unconsciousness" is clinical shorthand, not empirical fact.

The production model predicts that sufficient neural suppression should eliminate consciousness. The evidence shows it does not — at least not reliably. GNW and IIT can invoke auxiliary hypotheses (residual global ignition, isolated islands of high Φ), but as *Conscious Under Anesthesia* argues, both must explain why the evidence consistently runs opposite to their default predictions.

Under the constraint model, anesthesia disrupts the brain's capacity to *constrain and express* consciousness — not to produce it. Consciousness may persist in altered form even when the normal apparatus for structuring and reporting it is pharmacologically suppressed. The IFT data and the ketamine paradox are what this model would expect.

What the Tests Show

Neither the psychedelic nor the anesthesia evidence *refutes* any production model. As noted throughout, these models can invoke auxiliary hypotheses to accommodate the data. The question is not whether production models survive but whether they predict the patterns observed. The constraint model — which idealism enables — predicts both the psychedelic pattern (weakened constraints, expanded experience) and the anesthesia pattern (suppressed expression, persistent consciousness) without auxiliary hypotheses.

This is not decisive. But it establishes that idealism's interpretive framework generates a more parsimonious account of the empirical landscape than the production models do for these specific domains.

XI. Implications

For Consciousness Science

This analysis does not ask consciousness scientists to become idealists. It asks them to notice that their empirical findings are ontologically neutral and that the production inference is a commitment, not a discovery. The structural constraints identified by IIT, GNW, RPT, predictive processing, and HOT are genuine contributions regardless of ontology. Making the ontological packaging explicit allows the findings to be evaluated on their own terms — and applied within whichever interpretive framework proves most coherent.

In practice, this means:

- **The adversarial collaboration model is valuable but incomplete.** Testing IIT against GNW tests mechanistic predictions. It does not test — and cannot test — whether the brain produces or constrains consciousness. The deeper question requires a different kind of empirical investigation: one that tests the *direction* of the brain-consciousness relationship (production vs. constraint) rather than the *mechanism* within an assumed direction.
- **Anomalous findings deserve mechanistic attention.** Phenomena like terminal lucidity (coherent consciousness during severe neurodegeneration), near-death experiences during cardiac arrest, and persistent awareness under anesthesia are not merely curiosities. They are data points that test the production assumption. A science of consciousness that ignores them because they do not fit the default model is not being rigorous — it is being selective.
- **Ontological pluralism in research design is possible.** Experiments can be designed to test predictions that differentiate the production and constraint models — not as metaphysical advocacy but as scientific methodology. The psychedelic directionality pattern and the anesthesia persistence data are examples of evidence that carries differential pressure. More such tests can be devised.

For Philosophy of Mind

The landscape analysis reveals that the apparent competition among theories of consciousness is largely intramural — disagreements about mechanism *within* a shared physicalist assumption. The deeper question — whether consciousness is produced or fundamental — cuts across all these theories and is engaged by almost none of them explicitly. Making this visible allows philosophy of mind to distinguish two levels of question that are currently conflated:

1. **The structural question:** What are the features of consciousness? (Integration, accessibility, recurrence, prediction, unity.) This is answered by the theories and is ontologically neutral.
2. **The ontological question:** What is the fundamental nature of consciousness? (Produced by physical processes, fundamental and non-derived, an illusion, the intrinsic character of physical reality.) This is not answered by the theories, though most assume an answer.

For the Project

This essay fills a gap in the *Return to Consciousness* project. The existing essays establish the ontological neutrality of science (*The Generativity Question*), the historical contingency of physicalism's dominance (*The Emergence of Physicalism*), the asymmetry of skeptical standards (*Asymmetric Methodological Restraint*), the brute-fact structure of both frameworks (*Where Explanation Stops*), and the first-principles comparison (*First-Principles Assessment*). But they engage "physicalism" as a category rather than as a landscape of specific, well-developed theories.

This essay demonstrates that the project's diagnostic tools — constraint-based reasoning, ontological portability, the production/correlation distinction — apply not only to physicalism in the abstract but to the specific theories that constitute contemporary consciousness science. The result is not a conflict between idealism and the science but a distinction between the science (which is ontologically neutral) and its default packaging (which is physicalist but detachable).

XII. Limitations, Scope, and Honest Assessment

What this essay does not establish

This essay does not prove that idealism is correct. It does not refute any theory of consciousness. It does not show that consciousness science is flawed or misdirected. It does not claim that the structural findings of these theories are trivial once separated from their ontological packaging — they are genuine contributions to understanding consciousness regardless of ontological interpretation.

Where the analysis is strongest

The analysis is most secure in its diagnostic function: separating constraints from commitments, identifying the convergence pattern, and showing that the production inference is a commitment rather than a finding. These observations follow from the IBC methodology and the portability thesis established elsewhere in the project. They do not require accepting idealism.

Where the analysis is most exposed

Two vulnerabilities should be named plainly.

First: The structural-cost comparison (Section VIII) is a theoretical-virtue argument — a comparison of parsimony and explanatory coherence, not an empirical demonstration. This is the correct standard for ontological comparison, and physicalism's equivalent claims (that emergence handles consciousness, that neural correlates support production) operate at the same level. But a physicalist can respond that the apparent advantage is an artifact of framing — that once the right bridge principles are articulated, the derivation of mental features from non-mental processes will carry no more structural cost than idealism's account of why dissociation produces *these* specific minds. This response defers rather than dissolves the problem (the bridge principles remain unformulated), but it is not unreasonable.

Second: The idealist reinterpretation of each theory is relatively easy to state but has not yet generated the research programs and testable theories that would give it empirical traction beyond reinterpretation. As *The Generativity Question* argues, ontologies do not produce predictions directly — theories do, and theories are ontologically portable. The demand is therefore not that idealism derive predictions from its ontological axiom, but that researchers working within idealism's expanded space develop theories that generate novel, testable claims. The essay has identified two domains (psychedelics and anesthesia) where differential predictions between production and constraint models are already possible. Within this project, *Consciousness Structure* represents one such program: its boundary-coherence model operates within idealism's ontological space, generates six falsifiable predictions, and specifies operationalization pathways — demonstrating that idealism-native research is possible, not merely called for. But the broader development of such programs remains nascent.

The honest summary

Contemporary consciousness science has identified genuine structural features of consciousness: integration, global accessibility, self-reference, anticipatory modeling, and non-trivial unity. These features are ontologically neutral — they describe what consciousness *does* without settling what consciousness *is*. Most theories package these findings inside a physicalist production model that is assumed rather than entailed. Separating the findings from the packaging reveals that idealism accommodates the structural features with fewer unexplained transitions than physicalism — because the features are native properties of mind rather than emergent products of non-experiential processes. This is a theoretical-virtue comparison — the same kind of reasoning by which all ontological assessments proceed, including physicalism's own claim that emergence handles consciousness. Idealism's remaining task is not to derive mechanisms from its ontological axiom — no ontology does that — but to develop the research programs and testable theories that its expanded space makes possible. The question of which framework is ultimately more adequate can only be advanced by empirical investigations that test the production model against the constraint model — investigations that the current default ontology neither encourages nor funds.

Conclusion

The science of consciousness has achieved more than is commonly recognized — and less than is commonly claimed.

More: the structural features identified by IIT, GNW, RPT, predictive processing, HOT, and related theories are genuine discoveries about how consciousness is organized. Integration, accessibility, recurrence, prediction, and self-monitoring characterize conscious experience with growing empirical precision. These findings will endure regardless of which ontology prevails.

Less: these findings do not settle what consciousness is. The production inference — from correlation to generation — is a metaphysical commitment inherited from background physicalism, not a conclusion of the research programs themselves. Every theory examined here terminates at a version of the same hard problem: why does *this* process feel like anything? The specific processes differ; the gap between structure and experience persists.

Analytic idealism does not contest the findings. It offers a different account of what they describe. Under idealism, integration, self-reference, prediction, and global accessibility are not properties that physical processes must somehow generate from non-experiential substrates. They are properties of consciousness itself — features of what mind *does*, visible in the brain because the brain is the extrinsic appearance of how consciousness structures its own experience.

This is not the end of the inquiry. It is a reorientation. The question shifts from "How do physical processes produce consciousness?" — a question that has generated extraordinary mechanism-level work but remains stuck at the foundational level — to "How does consciousness structure itself in ways that appear as physical processes?" Both questions permit the same empirical investigations. But the second permits a wider space of conceivable theories, including directions the first forecloses.

The recurring structural findings of consciousness science describe features of mind. The ontology that takes mind as fundamental accommodates those features without the structural strain of deriving them from what lacks them. This does not prove idealism. It shows that the apparent conflict between consciousness science and consciousness-first metaphysics is not a conflict at all — it is an artifact of ontological packaging that, once made visible, can be set aside.

What remains is the science. And the science, honestly read, constrains every framework — including idealism, which must now generate the research programs and testable theories that its expanded ontological space makes possible. Some constraints sit equally within both frameworks; others, as Section X argues, create differential pressure. The work is not finished. But the space in which it proceeds is wider than the default assumption allows.

References

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-247.

Carhart-Harris, R. L., Erritzoe, D., Williams, T., et al. (2012). Neural correlates of the psychedelic state as determined by fMRI studies with psilocybin. *Proceedings of the National Academy of Sciences*, 109(6), 2138-2143.

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.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200-227.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

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.

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Consciousness*. Pantheon Books.

Graziano, M. S. A. (2013). *Consciousness and the Social Brain*. Oxford University Press.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews*, 11(1), 39-78.

Kastrup, B. (2019). *The Idea of the World: A Multi-Disciplinary Argument for the Mental Nature of Reality*. Iff Books.

Kelly, E. F., Kelly, E. W., Crabtree, A., Gauld, A., Grosso, M., & Greyson, B. (2007). *Irreducible Mind: Toward a Psychology for the 21st Century*. Rowman & Littlefield.

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

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

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

McFadden, J. (2020). Integrating information in the brain's EM field: The cemi field theory of consciousness. *Neuroscience of Consciousness*, 2020(1), niaa016.

Melloni, L., Mudrik, L., Pitts, M., et al. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642(8066), 133-142.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435-450.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press.

Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Dutton.

Shani, I. (2015). Cosmopsychism: A holistic approach to the metaphysics of experience. *Philosophical Papers*, 44(3), 389-437.

Stoljar, D. (2001). Two conceptions of the physical. *Philosophy and Phenomenological Research*, 62(2), 253-281.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10-11), 3-31.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450-461.

Tonetto, B. (2026). *Return to Consciousness: A Philosophical Journey from Materialism to Meaning*.

Related Essays in This Project

Available at: https://returntoconsciousness.org/

Integration by Constraints (ibc) — The methodology this essay applies

Where Explanation Stops (wes) — The brute-fact framework applied here to the full theory landscape

The Generativity Question (tgq) — The portability thesis and category error analysis

One Structure (ost) — Parallel analysis: convergence across contemplative traditions

Anomalous Phenomena and Consciousness (apc) — Detailed production vs. constraint comparison on empirical cases

First-Principles Assessment (fpa) — Where Russellian monism and panpsychism are engaged at first principles

Conscious Under Anesthesia (cua) — Anesthesia evidence discussed in Section X

License