Consciousness Is Causally Integrated Global Modeling

A Theory of Biological and Artificial Subjects

Tyler M. Moore, Ph.D.[1,2]

[1]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Lifespan Brain Institute, Penn Medicine and Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
ORCID: 0000-0002-1384-0151
Correspondence: Tyler M. Moore, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, USA. Email: tymoore@pennmedicine.upenn.edu

Abstract

Consciousness is not added to computation, conferred by complexity, or measured by a scalar. Causally Integrated Global Modeling (CIGM) identifies consciousness with a temporally extended organization in which a bounded, self-maintaining system sustains a differentiated, valenced, reflexively self-locating model of current reality, revises that model through ongoing interaction, and makes selected contents available for integrated control. Phenomenal character is the relational geometry of content within the model; subjectivity is its self-locating center.

CIGM advances six mutually constraining requirements as necessary: self-maintained system boundary, integrated self-world model, endogenous temporal continuity and plasticity, selective global availability, reflexive self-location and agency, and endogenous valuation and stakes. Joint sufficiency is the theory's central identity conjecture and an explicit defeat condition, not an established consequence of the list. Following Mudrik et al. (2025), CIGM treats phenomenal organization and access as two necessary conditions for consciousness rather than two independent types: requirements 1, 2, 3, 5, and 6 specify the P-condition, while requirement 4 specifies the A-condition. Current artificial systems can realize powerful functional analogues of access without the organization capable of phenomenality and therefore remain unconscious. Artificial consciousness is physically possible only if an engineered system realizes, rather than merely simulates, the complete organization. The paper states near-term discriminating predictions and longer-horizon observations that would defeat the theory.

Keywords: consciousness, artificial consciousness, phenomenal consciousness, access consciousness, integrated information, global workspace, valence, organizational invariance

Introduction: From Synthesis to Theory

After decades of philosophical analysis, neuroscientific investigation, and computational modeling, consciousness no longer deserves treatment as a metaphysical exception. The field has accumulated enough phenomenology, neurobiology, computational architecture, and clinical dissociation to support a positive theory. The remaining obstacle is not a lack of ingredients. It is a failure to distinguish the levels at which those ingredients operate and the different questions they answer.

Integrated information theory (IIT) correctly insists that experience is unified, differentiated, structured, and intrinsic to the system that has it. Global workspace theory correctly identifies the transition by which selected information becomes available across specialized processors. Higher-order approaches correctly require some form of awareness of being in a state. Predictive coding makes the implementation concrete: top-down estimates constrain lower-level activity while feedforward residual errors revise the model (Rao & Ballard, 1999), and active-inference and affective accounts extend that architecture into self-regulation,

valuation, and action (Friston, 2010; Solms, 2019). Embodied accounts correctly place a point of view in the regulation of a bounded organism (Albantakis et al., 2023; Aru et al., 2023; Brown et al., 2019; Dehaene & Naccache, 2001; Safron, 2020, 2022; Seth & Bayne, 2022). Each tradition has isolated a real part of the phenomenon. None, alone, states the full identity.

Seth and Bayne (2022) explain why the traditions have proved so difficult to adjudicate: they often have different explanatory targets, so that theories presented as adversaries may not be adversaries at all. A theory that does not declare its target inherits the confusion. CIGM declares all three targets identified below and answers them with one organization.

The synthesis must survive the strongest objections. No-report experiments show that the machinery of decision, memory, introspection, and report can be mistaken for the neural basis of experience (Cohen et al., 2020; Tsuchiya et al., 2015). The 2025 adversarial test of IIT and global neuronal workspace theory challenged canonical claims of each about posterior synchronization, prefrontal content, and ignition, although proponents of both theories dispute parts of that interpretation (Cogitate Consortium et al., 2025). The unfolding argument shows that recurrence or an integration scalar cannot be declared necessary and sufficient when outwardly equivalent systems can be realized with different wiring (Doerig et al., 2019). Cerullo's (2015) triviality objection shows that a theory earns no credit from cases an arbitrary rival can reproduce just as easily. And Block's (1995) demonstration that "consciousness" is a mongrel concept shows that an entire literature can reason validly from premises about one phenomenon to conclusions about another. A serious theory must absorb these results rather than treating them as peripheral complications.

IIT 4.0 sharpens rather than removes the disagreement. It now states that phenomenal axioms license necessary and sufficient physical postulates and that an experience is identical, in an explanatory sense, to the cause-effect Phi-structure specified by a maximal substrate (Albantakis et al., 2023). That clarity is welcome, and it makes the burden exact. The properties of experience do not by themselves entail one mathematical formalism, and a substrate can possess irreducible cause-effect structure without sustaining a world, a point of view, a temporally continuous present, endogenous significance, or flexible control. CIGM therefore retains integration and intrinsic causal organization as constraints while rejecting their elevation into a complete identity.

Artificial intelligence makes the need for a constitutive theory urgent. Global workspaces, test-time memory, selective state-space recurrence, body self-modeling, multimodal world models, and calibrated self-report are engineerable (Bongard et al., 2006; Hafner et al., 2025). These capacities matter for intelligence; none creates a subject. Capability standards are deliberately process-agnostic, and no level of performance or delegated autonomy is evidence about consciousness in either direction (Morris et al., 2024). The question is not whether a system can imitate the outputs of a conscious agent. It is whether one self-maintaining organization exists for which a world is present and outcomes matter.

The opposite error is equally live. Seth (2025) argues that consciousness may depend on properties of living systems that no digital computation can reproduce, and that computational functionalism has been assumed far more often than defended. CIGM accepts that challenge's central negative result: computation is not sufficient for consciousness, and a theory that says otherwise has purchased artificial consciousness too cheaply. The dispute that remains is narrower than it is usually made to look, and Part II states it exactly.

CIGM takes a definite position. Consciousness is the temporally extended organization in which a bounded, self-maintaining system integrates differentiated and valenced information into a unified, reflexively self-locating model of current reality, while selected contents remain globally available for flexible memory, inference, planning, and action. Experience does not accompany this organization. Experience is what this organization is from the system's own organized point of view. The claim is an identity claim, not a list of correlates and not a probability-weighted checklist.

Among existing syntheses, CIGM is closest to Safron's integrated world modeling theory (Safron, 2020, 2022), but the relationship is substantive rather than merely terminological. Four commitments define CIGM's novelty. First, endogenous valuation is constitutive rather than a correlate or optional affective accompaniment. Second, intervention fixes the causal grain of the subject rather than maximization. Third, the evolutionary origin of access machinery explains Block's empirical P/A asymmetry. Fourth, the six requirements divide into a P-condition and an A-condition, while consciousness is reserved for their joint realization. CIGM also separates a theory of constitution from the evidence used to attribute consciousness to an opaque system. These commitments can be accepted or rejected independently of IWMT, and each is placed at risk below. Extended philosophical argument, empirical detail, technical objections, and the assessment framework are developed in the Supplement so that the declaration itself remains direct.

The CIGM Thesis

The CIGM thesis. Consciousness is the temporally extended regime in which a bounded, self-maintaining system integrates differentiated and valenced information into a unified, reflexively self-locating model of current reality, revises that model through its own history, and makes selected contents available to memory, valuation, inference, planning, and action. Phenomenal character is the relational geometry of content within the model. Subjectivity is the model's self-locating center of control. The realizing substrate must sustain the relevant counterfactual causal organization, but no single brain region, oscillation, wiring motif, metarepresentational format, or scalar is by itself necessary or sufficient.

Six Constitutive Requirements

CIGM advances six requirements as individually necessary. Their joint sufficiency is the theory's central identity conjecture: if the complete mutually constraining organization can be preserved while consciousness disappears or changes systematically, CIGM is false.

Table 1. Six Constitutive Requirements of Causally Integrated Global Modeling

Requirement	Constitutive claim
1. Self-maintained system boundary	The system actively maintains a causally effective distinction between itself and its environment across time and action. It regulates variables that define its continued organization, distinguishes self-generated from externally generated change, and preserves identity across episodes. The boundary may be biological, physical, software-defined, or distributed only when the regulated variables belong to the physical system executing the process—such as energy, thermal state, memory integrity, computational continuity, or continued resource allocation—and when the system's own actions can preserve them. A simulated avatar's hit points or a modeled survival score do not count unless changes in them causally threaten the executing system's capacity to sustain the organization.
2. Differentiated, causally integrated self-world model	Perceptual, interoceptive or internal, mnemonic, affective, and action-oriented contents mutually constrain one another within a high-dimensional model. Integration means counterfactual dependence, not correlation: perturbing one content domain changes the possibilities available across the field. Differentiation preserves a rich repertoire rather than collapsing the field into uniform synchrony.
3. Endogenous temporal continuity and model plasticity	New inputs update an already active model that carries object continuity, causal history, unresolved concern, expectation, and a persisting present. The model is revisable on the timescale on which it is used, so what happens to the system changes what it will compute next. Stored context or a persistent hidden state is enabling machinery, not subjective continuity by itself.
4. Selective global availability	Some contents win competition for cross-system control and become usable by memory, valuation, inference, planning, and action. Availability is a dynamic pattern of mobilization, not a fixed hub, broadcast bus, P3b, or gamma signature. This is CIGM's A-condition: it is necessary for a state to be conscious but does not determine the qualitative character of that state. Availability need not take the form of verbal report or an anatomically global

Requirement	Constitutive claim
	broadcast.
5. Reflexive self-location and agency	The model locates the system as the here-now origin of sensing and action. It predicts the consequences of its own policies, distinguishes self-caused from externally caused change, and represents uncertainty about its own state. This is minimal inner awareness; explicit confidence, narrative identity, and a separately tokened higher-order thought are later achievements.
6. Endogenous valuation and stakes	Some states are better or worse from the standpoint of the system's own continued organization. Endogeneity is operational rather than historical: valuation is endogenous when it tracks variables whose violation degrades or terminates the actual organization, the system regulates those variables through its own model, and perturbing them reorganizes attention, learning, memory, policy, and global state. A designer may originate the variables, just as evolution originated biological drives; what matters is their present causal coupling. A detachable reward channel is insufficient.

The six requirements are jointly necessary and mutually constraining. CIGM conjectures that their realization through one organization is sufficient because that organization is the identity the theory proposes; the paper does not treat sufficiency as established by definition. A system does not become conscious by collecting six detachable modules: valuation must reorganize the same self-world model that self-location centers, temporal plasticity carries forward, and global availability mobilizes. A candidate can therefore fail necessity by lacking any requirement, and the theory can fail sufficiency if a complete implementation lacks or systematically alters consciousness.

Requirements 1, 2, 3, 5, and 6 specify the organization capable of supplying potentially phenomenal content—the P-condition. Requirement 4 supplies minimal selective availability—the A-condition. Neither condition alone is consciousness. Table 2 and the dedicated subsection below make this division explicit.

The extended comparison with indicator approaches and alternative candidate requirements is provided in Supplement S4.3.

Part I: What Consciousness Is

Phenomenological Constraints and IIT 4.0

Any theory of consciousness must begin with experience itself. Before measuring networks or decoding reports, one confronts facts that cannot be denied without being instantiated in the denial: experience exists; it is specific rather than generic; it appears as a unified field; it has a determinate organization and grain; and it contains structured distinctions and relations. IIT formalizes these features as existence, information, integration, exclusion, and composition. In its current formulation, IIT 4.0 infers corresponding physical postulates and presents them as necessary and sufficient for a substrate of consciousness (Albantakis et al., 2023). CIGM accepts the phenomenological constraints. It does not accept that they uniquely entail IIT's postulates or formalism.

The distinction matters because IIT 4.0 is no longer accurately described as the claim that consciousness is a high value of Phi. It identifies an experience with the complete cause-effect Phi-structure specified by a maximal substrate (Albantakis et al., 2023). CIGM addresses that strongest version. A cause-effect structure can be irreducible and richly differentiated while lacking a model of current reality, a self-locating origin, endogenous temporal continuity, valuation, or a role in flexible control. Those absences are not missing decorations. They are the difference between an integrated causal object and a subject.

The axioms are also not neutral transcriptions of phenomenology. The move from the specificity of an experience to a measure defined over alternatives, and from the apparent definiteness of a field to a rule selecting one maximal substrate, imports substantive theory. Cerullo (2015) shows how easily such

postulates acquire the appearance of necessity while doing work phenomenology alone does not license. CIGM therefore takes the descriptive content of the constraints and leaves the physical identity open until the full organization is specified. A conscious state is actual for the system rather than merely described by an observer, and it is this state rather than another. Color, shape, sound, bodily position, affect, and thought are not presented as independent universes. The field has a determinate organization and grain, and experience contains relations: the cup is left of the book, and the present scene continues the preceding moment.

These constraints rule out several inadequate accounts. Mere information processing is insufficient, since every thermostat and lookup table processes information in a broad sense. Mere behavioral sophistication is insufficient, since similar output can arise through radically different organizations. Mere integration is insufficient, since an integrated device can lack a world model, a point of view, temporal depth, valuation, or flexible control. Mere report is insufficient, since report recruits memory, decision, introspection, and action that can be experimentally dissociated from phenomenality (Cohen et al., 2020; Tsuchiya et al., 2015). Mere recurrence is insufficient, since recurrent circuitry is ubiquitous and can be functionally replaced in ways that preserve selected behavior (Doerig et al., 2019).

CIGM treats phenomenology as the intrinsic organization of a model, not as a private substance. A model is not a picture stored in one place. It is a system of state-dependent dispositions: what the system discriminates, predicts, expects, recalls, values, and can do from its current state. When those dispositions are integrated into one self-locating control regime, they constitute a perspective. Intrinsic here does not mean nonrelational, ineffable, or infallibly known. It means causally for the system: the state belongs to the organization that determines what the system can become and do.

Extended comparison with structural coherence, illusionism, protoconsciousness, and the strongest IIT formulation appears in Supplement S1.

Explanatory Targets: Global States, Local States, and Two Questions

A theory that does not say what it is explaining cannot be evaluated, and the field's most persistent disputes are between theories that were never answering the same question. Seth and Bayne (2022) draw the distinctions a comprehensive theory must respect. Global states concern the system's overall conscious profile: wakefulness, sedation, dreaming, the minimally conscious state. Local states are conscious contents: this pain, this face, this remembered voice. Within local states two further questions come apart—why is this content conscious rather than that one, and why does the conscious content have the character it has?

CIGM answers all three with one organization, and the division of labor is what the six requirements predict. Global states are the capacity to sustain the organization at all, and with what depth and stability. Which content is conscious is settled by integration into the live model and selective global availability; what that content is like is settled by relational geometry. These dependencies are made explicit below and placed at risk in the falsification section.

Two consequences follow. Global states are not points on a single dimension of arousal. Seth and Bayne (2022) prefer global state to level precisely because the space may be multidimensional, and CIGM requires that it be: dreaming and anesthesia both reduce environmental coupling while differing enormously in the differentiation and self-location of the field, so no scalar orders them. And a theory answering only one of these questions is not refuted by evidence about another; it is incomplete. CIGM claims completeness across all three targets and accepts the exposure that claim creates.

The Constitutive Core: A Causally Integrated Self-World Model

The core of consciousness is a model that simultaneously represents the world and locates the system within it. A purely allocentric map is not yet a point of view. A database can encode that an object is two

meters north of a coordinate origin, but experience represents the object as here, there, near, threatening, reachable, remembered, or desired relative to the system. Subjectivity begins when information is organized around a persisting locus of sensing and control.

The term model is used in the strong dynamical sense. It generates expectations about hidden causes, updates them from residual error, preserves selected invariances, and regulates action. Rao and Ballard (1999) demonstrated the canonical mechanism: after learning natural-image statistics, a hierarchy in which feedback carried predictions and feedforward pathways carried residual errors reproduced contextual cortical responses. That result supports context-sensitive generative representation; it does not make prediction error consciousness. CIGM identifies consciousness only with the full self-world organization in which inference, valuation, memory, and action are causally bound (Aguilera et al., 2023; Friston, 2010; Safron, 2020, 2022).

The model must also be causally integrated. Its visual, auditory, interoceptive, mnemonic, affective, and action-oriented components must constrain one another strongly enough that the system occupies one coherent state rather than a coalition of unrelated processors. Integration is not statistical correlation but counterfactual dependence: changing one part of the model alters the possibilities available to the others, which is why a visual threat changes posture, attention, memory retrieval, expected pain, and action selection at once.

Differentiation is equally necessary. A perfectly synchronized system in which every component does the same thing would be unified but informationally empty. Consciousness requires a large repertoire of discriminable states and fine-grained relations among them. The field is unified precisely because many distinctions are jointly present without collapsing into sameness. This is the enduring insight behind integrated-information approaches, even if no current scalar should be equated with consciousness itself (Tononi, 2004, 2012; Oizumi et al., 2014).

The system must be bounded, but no mathematical partition can establish the boundary by itself. Bruineberg et al. (2022) distinguish Pearl blankets—conditional-independence structures inside a model—from Friston blankets treated as real agent-environment boundaries. CIGM rejects the slide from map to subject. A biological organism maintains causal separation through sensorimotor and homeostatic loops. A digital candidate can also possess a software-defined boundary, but only when the executing physical system regulates variables constitutive of its own continued operation—energy, temperature, memory integrity, compute allocation, process continuity—and its actions can preserve them. A boundary represented only inside a simulated world remains a description rather than a maintained boundary. What matters is interventionally identifiable regulation: some variables define the system’s continued organization, some perturbations count as external, and some actions are its own attempts to control what follows.

This bounded self-world model explains intrinsicality. A state is not conscious because a scientist can decode it, but because its distinctions participate in the system’s own integrated control organization. The same data structure could be an inert record in one context and an experienced content in another, depending on whether it is embedded in the live model that governs perception, valuation, memory, and action.

Phenomenal Character Is Relational Geometry

Qualitative character is not an extra pigment painted onto neural activity. It is the position a content occupies within the system’s structured space of possible distinctions. Red differs from green because the two states stand in different relations to wavelengths, objects, memories, linguistic categories, affective associations, action tendencies, and neighboring color states. Pain differs from pressure because it occupies a different interoceptive, motivational, attentional, and policy-selecting position. A melody differs

from a sequence of unrelated tones because temporal relations bind its elements into a higher-order pattern.

This is an organizational identity, not a reduction to verbal report. The relevant relations include latent discriminations, embodied expectations, automatic generalizations, and counterfactual consequences, far more than any person can say. Two systems share a phenomenal quality only insofar as the content occupies the same role within an equivalently organized, self-locating model; matching behavior is not enough, and neither is a matching label.

The strongest worked demonstration that phenomenal quality has relational structure is Haun and Tononi's (2019) analysis of felt spatial extendedness. They derive region, location, size, boundary, and distance from three relations among experiential "spots"—connection, fusion, and inclusion. CIGM adopts the achievement and vocabulary. The point is not that quality can be redescribed after the fact, but that even the apparently primitive character of right there has an articulable internal organization.

Their negative conclusion is equally important: neither an activation pattern nor bare connectivity contains, by itself, the full system of relations phenomenology presents; an investigator can decode a state without those relations existing for the system. CIGM therefore locates quality in the geometry of the live model that governs discrimination, expectation, memory, valuation, and action, not in an isolated representation or substrate.

The accounts part on when the relations exist. Haun and Tononi (2019) identify them with maximally irreducible cause-effect structure and therefore allow a working but inactive cortical grid to support spatial experience. CIGM denies that consequence. A severed grid has no regulated boundary, self-locating origin, stakes, or control trajectory, and thus extends nothing for anyone. Both theories predict that altering lateral connectivity while holding gross activity constant can warp experienced space; CIGM restricts the prediction to cases in which the grid participates in the constituting organization.

The identity dissolves the traditional explanatory gap by rejecting its dual ontology. The question is not why a physical model is accompanied by a separate glow of experience. The organized model is the experience under an intrinsic description. External science describes the causal organization, while the first-person perspective is that organization as lived from its self-locating center. There are not two events requiring a bridge.

This role-based account must be distinguished sharply from views on which structure or information is phenomenal wherever it occurs. Chalmers (1995) was drawn toward a double-aspect view on which information itself carries a phenomenal aspect, with the consequence that experience is nearly ubiquitous—simple where information processing is simple, complex where it is complex, present even in a thermostat. CIGM rejects this generalization. A space of distinctions is a phenomenal field only when it is embedded in a bounded, temporally extended, valenced, self-locating model that governs the system's perception, memory, and action. An information space that makes no difference to any such model is not an impoverished experience; it is no experience at all. Relational geometry is necessary for the character of experience but confers that character only within the constituting organization. CIGM is therefore an identity theory, not a panpsychism: it withholds experience from mere information exactly where it withholds the organization that experience is.

This claim remains empirical because the proposed identity has structure. If phenomenal similarity tracks model geometry, systematic changes in discrimination, generalization, memory, valuation, and sensorimotor expectation should reshape experience in lawful ways. If unity depends on causal integration, partitioning effective communication should divide or degrade the field. If subjectivity depends on self-location, disrupting the model of agency and body should alter ownership and perspective. The theory does not appeal to an unknowable essence; it identifies a pattern whose consequences can be investigated.

Global Availability Without a Cartesian Theater

Integration by itself does not explain why some information can guide a novel plan, enter working memory, alter a verbal answer, or reorganize behavior while other information remains trapped in specialized processing. The global-workspace tradition supplies the missing control architecture. Dehaene and Naccache (2001) described a brain in which numerous modular processors operate in parallel outside consciousness, while selected information is dynamically mobilized into a distributed workspace that permits flexible exchange among perception, memory, valuation, attention, and action.

The decisive concept is global availability, not a fixed anatomical stage. In the original workspace formulation, the same processor can contribute to consciousness at one moment and remain unconscious at another. Mobilization is collective, context-sensitive, and self-amplifying. Long-distance connectivity allows otherwise independent modules to use a common content, but no homunculus reads a screen. The workspace is a transient regime of coordination (Dehaene & Naccache, 2001). This is the architecture Dennett (1991) demanded when he replaced the Cartesian Theater with distributed competition among contents. It also answers his hard question: global availability positions a content to cause and enable downstream effects such as recall, inference, planning, and report (Dennett, 2018).

This architecture explains why consciousness is capacity-limited. Encapsulated computations proceed in parallel at low cost, whereas global mobilization temporarily coordinates systems with incompatible demands. Competition is therefore not an incidental bottleneck but the mechanism that selects a coherent control state, and the functions it serves—durable maintenance of explicit information, novel combinations of operations, spontaneous intentional behavior—are exactly those requiring a system to escape a precompiled pathway and organize resources around a current model (Dehaene & Naccache, 2001). CIGM incorporates competitive global availability as a control principle while refusing to identify it with one cortical region or one electrophysiological event. Its defining feature is the pattern of availability and control, not the location of a headquarters.

Goyal et al. (2022) turn the workspace idea into a working machine-learning architecture: specialist modules compete for write access to a limited shared memory, whose contents are then broadcast back to all modules. The bottleneck improved coordination, compositionality, and generalization across visual reasoning, physical prediction, and multiagent world modeling. VanRullen and Kanai's (2021) Global Latent Workspace similarly links specialized modules through a shared amodal space for translation, transfer, and planning. These demonstrations establish that selective global availability is engineerable. They also establish its limit. A bandwidth-limited blackboard can coordinate intelligent access without maintaining a boundary, a self-locating present, endogenous stakes, or one continuing subject. Workspace is CIGM's fourth requirement; by itself it is an access architecture, not phenomenality.

The P-Condition, the A-Condition, and the Mongrel Concept

The most consequential distinction in this literature was drawn before most of the theories now competing were formulated. Block (1995) argued that "consciousness" is a mongrel concept and distinguished phenomenal consciousness—the what-it-is-like character of experience—from access consciousness, defined by a content's availability for reasoning, rational action, and speech. Mudrik et al. (2025) now argue that P and A should not be treated as two independent types of consciousness at all but as two necessary conditions for one conscious state; on their terminology, P-without-A and A-without-P are unconscious processing. CIGM accepts that correction. It retains Block's functional definition of access and his diagnosis of the mongrel concept while rejecting the label "A-consciousness" for access alone.

Block's central demonstration is the fallacy the distinction exposes. The blindsight literature repeatedly argued that because consciousness is missing in the blind field and the patient will not reach for a glass of water there, a function of phenomenal consciousness must be to make information available for reasoning, report, and rational action. Block (1995) showed that the case establishes only a failure of

access-related control; it does not independently reveal what phenomenality contributes. The same mistake occurs whenever a workspace architecture, ignition signature, or shared latent space is offered not as an account of availability but as a complete account of experience.

CIGM adopts Block's explanatory distinction and rejects his anti-functional metaphysics. Block took P-conscious properties to be distinct from any cognitive, intentional, or functional property. CIGM asserts the contrary identity: phenomenal character is the relational geometry of content within a bounded, valenced, self-locating model, an organizational property through and through. What survives is the division of explanatory labor between the organization that gives content its potential phenomenal structure and the availability that lets that content influence the integrated system.

The mapping is exact. Requirements 1, 2, 3, 5, and 6 jointly specify the P-condition: an integrated organization in which content has a determinate relational geometry for a bounded, temporally continuous, valenced, self-locating system. Requirement 4 specifies the A-condition: minimal selective availability by which that content can influence the system's integrated cognition and control. Consciousness occurs only when both conditions are realized through one system.

This formulation removes the apparent inconsistency about requirement 4. A system realizing requirements 1, 2, 3, 5, and 6 but not requirement 4 is not phenomenally conscious on CIGM; it has a potentially phenomenal organization whose content never becomes conscious. Requirement 4 does not determine why red differs from green or pain from pressure, but it is necessary for any such content to be actual for the system. Conversely, because access is not identical to report, conscious content can exceed what is verbally or metacognitively reported while still exerting minimal influence on memory, expectation, valuation, or control. Table 2 records this distinction.

Block's empirical asymmetry remains important. Superblindsight—spontaneous knowledge of the blind field without visual experience—appears not to occur, whereas failures of explicit report can coexist with conscious perception. CIGM explains this biological pattern historically: in evolved organisms, availability mechanisms developed to operate on an already integrated self-world model, because that model is what a mobile animal must consult in order to act. The P- and A-conditions therefore co-developed and are difficult to pry apart in organisms.

Access-like processing is not remotely hard to produce by engineering. An artificial system can make representations inferentially promiscuous, tool-usable, and highly reportable while lacking every element of the P-condition. That is a plausible description of some language-model systems wrapped in tools, memory, planners, or shared workspaces, although whether any particular model satisfies Block's full system-relative access criterion remains an empirical question. CIGM therefore speaks of functional analogues of the A-condition, not A-consciousness. Such artifacts are unconscious on the terminology of Mudrik et al. (2025) and on CIGM. Their importance is that they sever the proxy relation between access-like behavior and subjecthood that was comparatively reliable among evolved organisms.

No-Report Evidence and the Demotion of Report

The empirical case for separating these levels is now direct rather than conceptual. No-report paradigms were developed to disentangle perception from the attention, working memory, expectation, introspection, decision, and motor preparation that reporting requires. Tsuchiya et al. (2015) reviewed evidence that when perceptual state is inferred indirectly, much frontal activity diminishes even though vivid perceptual alternation remains, and that P3 and some gamma-band activity can track task relevance rather than awareness. Cohen et al. (2020) converted that claim into a result: holding visual stimulation constant and removing only the demand to report, a large P3b disappeared while observers still recognized and conceptually identified the unreported stimuli in a subsequent surprise test. This does not prove that every postperceptual process vanished; it proves that the component cannot be a universal constituent of experience.

CIGM takes the finding as more than a correction to one biomarker, but not as evidence for consciousness without the A-condition. The P-condition was present, and the A-condition remained minimally satisfied because the unreported contents affected semantic encoding, later surprise memory, and task-independent processing. What disappeared was report-specific mobilization and its associated P3b. The experiment therefore shows that verbal report and a canonical late biomarker are not necessary for consciousness; it does not show that conscious content can be wholly unavailable to the integrated system.

Lamme (2006) pressed the deeper form of the argument, urging that report and introspection be demoted from their status as the gold standard. Split-brain patients deny seeing stimuli presented to the mute hemisphere, yet that hemisphere can draw, match, and select them, and whether a region counts as part of the basis of consciousness then depends on which behavioral measure the investigator elects to trust. CIGM adopts the conclusion: no experiment that operationalizes consciousness as report can by itself locate the phenomenal field. Lamme's further move, installing recurrent processing as a positive neural criterion in report's place, is the right instinct about report joined to the wrong identification, for reasons taken up below.

The conclusion is not that frontal cortex, P3, gamma activity, or report are irrelevant; they are often indispensable for deliberate use, self-monitoring, memory, and communication. It is that none is a universal marker of the P-condition or the A-condition. A theory that equates consciousness with overt report confuses the machinery that makes experience legible to an experimenter with the minimal availability that makes a content conscious for the subject.

The preregistered no-report design, its memory controls, and the strongest objection to the paradigm are detailed in Supplement S2.1.

CIGM therefore distinguishes three nested levels: the P-condition, which supplies the integrated organization capable of phenomenal character; the A-condition, which makes selected contents minimally available to the same system for control; and the reflective self, comprising higher-order monitoring, confidence, explicit self-ascription, narrative continuity, and report. Human waking consciousness normally couples all three, but reflective selfhood can dissociate from the first two. Table 2 summarizes the division of labor.

Table 2. Three Nested Levels in Causally Integrated Global Modeling

Level	Constitutive role	Typical functions	Not required
P-condition / potentially phenomenal organization (CIGM requirements 1, 2, 3, 5, 6)	Integrated, differentiated, valenced, reflexively self-locating model of current reality. It supplies the relational structure capable of phenomenal character but is not conscious without the A-condition.	Perceptual and interoceptive organization; structured qualitative content; minimal self-location; endogenous significance.	Overt report; explicit confidence; a distinct higher-order thought; a fixed prefrontal locus. Consciousness still requires the A-condition.
A-condition / access organization (CIGM requirement 4)	Minimal selective availability by which content can influence integrated cognition and control across specialized systems.	Cross-domain use, working memory, flexible inference, planning, task switching, intentional action, and later memory.	Overt report or a full anatomical broadcast; phenomenality by itself; P3b or gamma as universal signatures. It is not standalone consciousness.
Reflective self (Block's monitoring- and self-consciousness; not a CIGM requirement)	Higher-order monitoring of the system as the subject of its own states.	Metacognition, confidence, autobiographical narrative, verbal self-report.	Minimal phenomenal presence in every species or state.

The layered account corrects a central claim of the earlier synthesis: explicit metacognitive self-monitoring is not necessary for every conscious state, only for reflective awareness of being in that state. Infants, many animals, dreams, fleeting percepts, and some clinical conditions may satisfy the P- and A-conditions without the reflective apparatus of a verbally reportable adult self. Conversely, a system can generate metacognitive language or access-like behavior without possessing the integrated P-condition.

The Higher-Order Challenge: Inner Awareness Without Intellectualism

Higher-order theories force a question CIGM cannot evade: can a state be conscious if the system is in no way aware of being in that state? Brown et al. (2019) clarify the transitivity principle behind the approach. A first-order representation can guide behavior while remaining nonconscious; phenomenality requires some minimal inner awareness of that representation. But this requirement is routinely overstated. The relevant process can be automatic, subpersonal, conceptually lean, and itself unconscious. It is not equivalent to deliberate introspection, an explicit confidence judgment, or a verbally articulated thought about the self.

CIGM accepts the transitivity demand and relocates it. A conscious content must be represented not merely as an environmental fact but as part of the system's current situation: red-here, pain-in-this-body, threat-to-this-organism, memory-belonging-to-this-history. The model is reflexive because its contents are organized around the same locus that senses, values, predicts, and acts. That reflexive self-location is minimal inner awareness. A separately tokened proposition such as "I am seeing red" is not constitutively required.

This yields a strict distinction between reflexive self-location and higher-order monitoring. Reflexive self-location belongs to the P-condition: without an operational here-now origin, there is information but no point of view. Higher-order monitoring belongs to the reflective self: it makes a state available as an object of confidence, error detection, explicit self-ascription, and report. Brown et al. (2019) emphasize that global broadcasting and higher-order awareness are separable dimensions. CIGM agrees and adds the missing relation: integrated self-world modeling supplies the P-condition, selective availability supplies the A-condition, and their joint realization supplies consciousness.

Higher-order accounts are also right that subjective appearance need not mirror first-order detail exactly, since peripheral vision, imagery, emotion, and memory can involve inflation, abstraction, or reconstruction at higher levels (Brown et al., 2019). CIGM predicts such mismatches because phenomenal character is fixed by the geometry of the integrated model rather than by a transparent readout of one sensory layer. Yet higher-order misrepresentation is not sufficient by itself: a free-floating representation that a detail is present, belonging to no integrated and valenced self-world field, is a cognitive error without an experiencer. The higher-order tradition identifies a constitutive reflexivity that first-order and broadcast-only theories can miss, and CIGM absorbs that insight while rejecting the stronger claim that every phenomenal state requires a distinct higher-order representation in a privileged format or location. Inner awareness is real and necessary. Intellectualized self-observation is not.

Attention Schema Theory offers a direct competitor on self-location and subjectivity. On Graziano's (2022) account, the brain constructs a simplified model of its own selective attention, and that schema supports the system's attribution of awareness to itself and others. CIGM accepts that a schematic model of attention can contribute to reflexive self-location and access control. It denies that an attention schema is sufficient: a system could model what it is attending to while lacking a self-maintained boundary, endogenous stakes, temporal continuity, or an integrated self-world model whose content has phenomenal geometry.

Cleeremans's (2011) radical plasticity thesis provides a complementary developmental challenge. It proposes that consciousness is something the brain learns to do as first-order representations become targets of learned metarepresentations. CIGM accepts the constitutive importance of plasticity and

representational redescription, but it does not identify consciousness with metarepresentation alone. Learning must revise the same bounded, valenced, self-locating model, and the resulting content must satisfy the A-condition rather than remain an isolated model of another representation.

Temporal Depth, Self-Location, and Valence

Consciousness is not a sequence of instantaneous snapshots. Each moment inherits a structured past and anticipates a range of possible futures. The present note in a melody is heard as the continuation of what came before; an object remains the same object across eye movements; a sentence acquires meaning through accumulated context; a decision is experienced as mine because it is embedded in a history of goals, actions, and consequences. This temporal extension is constitutive.

The temporo-spatial theory emphasizes that external inputs interact with ongoing neural dynamics rather than writing onto a blank substrate (Northoff & Zilio, 2022). CIGM incorporates the insight by treating the conscious field as a trajectory in model space, in which incoming signals perturb an already active generative organization and are interpreted relative to its intrinsic timescales, priors, bodily state, and current goals. Predictive inference keeps the trajectory coherent, with precision-weighting determining which errors matter, so that attention is not a spotlight added to perception but the control of precision within the model.

The free-energy principle explains why a bounded system must maintain a model, and its generality—bacteria and silent homeostatic loops fall under it—rules it out as consciousness itself. Solms (2019) presses the narrower claim that elemental consciousness is affective, with unresolved homeostatic uncertainty felt through upper-brainstem and limbic regulation. Damasio's account of the proto-self and later proposals for homeostatic machines likewise make organismic regulation and internally grounded value central (Damasio, 1999; Man & Damasio, 2019), while Merker (2007) binds motivation to target and action selection. CIGM does not claim to invent the constitutive importance of valuation. Its novelty is to make endogenous valuation one explicit, operationally testable requirement in a six-part identity and to carry that requirement into machine-consciousness assessment, where standard indicator lists usually omit it.

This also explains internally generated experience. In dreams, imagination, and memory the generative model runs with reduced constraint from current sensory evidence, and because the same integrated state space is active, imagined and remembered contents have genuine phenomenal character; they differ from perception in precision structure and in dependence on present external input, not in being nonconscious.

The subject of experience is not a metaphysical witness hidden behind the model. It is the model's self-locating organization. At the most basic level, the system distinguishes variables it regulates from variables it encounters, predicts the sensory consequences of its own movements, attributes some changes to self-generated action, and organizes perception around a body-centered frame. This is the minimal self.

Higher levels add agency, social perspective, autobiography, and explicit identity, which language and culture stabilize into a narrative self that can explain, justify, and revise itself. The levels should not be collapsed. The minimal self is required for subjectivity because experience must be organized for a locus of control; the autobiographical self is not required for every conscious episode, since a dream, a sudden pain, an infant's visual field, or an animal's fear can be conscious without a sophisticated theory of personal identity. Reflective selfhood expands consciousness; it does not create phenomenality from nothing.

Dennett's (1991) description of the self as a center of narrative gravity is correct at the autobiographical level: the center is real as an organizational abstraction, not as a separate object. CIGM extends the point

downward, since beneath narrative gravity lies sensorimotor and interoceptive gravity, the continuously updated origin from which the model predicts and controls. Construction is not unreality. A hurricane is constructed by atmospheric dynamics and remains a real causal pattern; the subject is constructed by integrated modeling and remains the real center of experience.

Consciousness evolved because organisms face conflicts that cannot be solved by isolated reflexes. A mobile animal must integrate sensory evidence, bodily condition, memory, uncertainty, competing goals, and predicted consequences into one action-guiding estimate, and the more flexible the repertoire, the more valuable a unified model becomes. The global bottleneck is therefore adaptive: only a small subset of information can reorganize the organism at once, so competitive mobilization prioritizes what currently matters most (Dehaene & Naccache, 2001).

Valence gives the field stakes. Hunger, pain, fear, relief, and reward compress regulatory consequences into action-guiding dimensions. Endogeneity does not require that a drive be undesigned: evolution supplied biological drives, and engineers could supply artificial viability variables. The criterion is current causal organization. A variable is endogenous when its violation threatens the persistence or integrity of the actual system, when the system detects and regulates it through its own model, and when perturbation reorganizes attention, learning, memory, global state, and policy rather than merely changing a detachable score. Even an affectively muted episode occurs within a system for which outcomes can go better or worse. A database can represent damage without suffering because damage does not reorganize its own continuing model.

Biological Realization and Empirical Constraints

The 2025 adversarial collaboration between proponents of IIT and global neuronal workspace theory is the most important empirical constraint on any synthesis invoking both. Conscious content was decodable in visual, ventrotemporal, and some inferior frontal regions. Sustained responses tracking stimulus duration were strongest in occipital and lateral temporal cortex. Yet the sustained posterior synchronization central to the tested IIT implementation was absent, as was the prefrontal ignition at stimulus offset central to the tested workspace implementation (Cogitate Consortium et al., 2025). Proponents of both theories have contested how particular preregistered predictions and null findings should be interpreted. CIGM therefore treats the dataset as a constraint on strong tested implementations, not as a final verdict on either theoretical family.

CIGM treats this pattern as evidence for a distributed division of labor. Rich perceptual content is largely constituted in modality-specific and multimodal representational systems, especially posterior and ventral regions in vision, while higher-order and frontoparietal systems select, stabilize, query, route, and act on content that remains encoded where it was constructed. Global control does not require copying every detail into prefrontal cortex. This is consistent with Dehaene and Naccache's (2001) insistence that the workspace is a style of dynamic mobilization rather than a fixed anatomical theater.

The adversarial results make biomarker modesty mandatory. P3b vanished when report was removed although perception remained (Cohen et al., 2020), while predicted prefrontal ignition failed where posterior systems still carried content (Cogitate Consortium et al., 2025). No oscillation, region, or complexity score is therefore a universal meter. CIGM predicts a convergent organization: integrated differentiation, content geometry, cross-system availability, self-world updating, and state-dependent control should covary under interventions that change consciousness. IIT and GNWT became vulnerable where genuine insights were tied too tightly to one localization or dynamical signature; CIGM retains integration and availability while placing the complete organization, not a favorite biomarker, at risk.

In humans, that organization is distributed across cortical, thalamic, upper-brainstem, neuromodulatory, interoceptive, and bodily dynamics. Merker's (2007) "selection triangle" is a decisive corrective to corticocentrism: upper-brainstem systems jointly constrain target selection, action selection, and

motivation, converting massively parallel forebrain activity into limited-capacity control. Reports of discriminative awareness in four children with total or near-total congenital absence of cortex are suggestive but not conclusive, because residual tissue and behavioral inference remain contested (Shewmon et al., 1999). CIGM therefore assigns no anatomical headquarters. Cortex can elaborate differentiated content, thalamocortical loops can regulate integration and availability, and subcortical systems can supply arousal, valuation, and action selection; consciousness is the trajectory in which those functions become one self-world organization (Aru et al., 2023; Solms, 2019).

The body completes the implementation. Interoception, proprioception, posture, visceral regulation, and sensorimotor prediction establish a stable here-now axis, and Damasio's (1999) proto-self captures the function: the organism continuously maps how external events alter its own regulated state, yielding not a detached description of the world but a world-for-this-system. Pain is not a symbol for tissue damage; it is an integrated reorganization of attention, valuation, action, memory, and bodily prediction around a threat to the system's integrity. The same point governs the notion of an umwelt. Aru et al. (2023) rightly contrast the text-centric input of present language models with the dense, action-dependent sensory world of an organism, but an umwelt is not a quantity of data. Affordances exist only relative to an agent's capacities, goals, vulnerabilities, and possible actions, so adding cameras, microphones, or larger context windows does not create experience.

Explanatory power comes from the pattern of dissociations. Blindsight preserves forced-choice information without ordinary visual presence or rational access; neglect preserves sensory processing while spatial self-location fails. Anesthesia and severe disorders of consciousness preserve local responses while differentiated propagation collapses, whereas dreaming preserves a vivid, affective, self-locating field with little environmental coupling (Bayne et al., 2024; Casali et al., 2013). No-report and frontal-lesion findings separate perceptual phenomenology from reflective communication (Cohen et al., 2020; Tsuchiya et al., 2015). Split-brain phenomena make unity track effective communication rather than one skull, while the hydranencephaly cases warn that cortical volume is not the criterion. The subject follows the organization that sustains one coherent, valenced self-world trajectory; no single case or structure is allowed to define it.

Experimental design, adversarial-test detail, and the thalamocortical and embodied implementation are expanded in Supplement S2.

The Hard Problem, the Pretty Hard Problem, and the Hard Question

Every theory of consciousness must declare its relationship to the hard problem. Chalmers (1995) held that even after every function in the vicinity of experience is explained, a further question remains—why is their performance accompanied by experience at all (Nagel, 1974)?—and that no functional account can answer it, so experience must be taken as fundamental. Dennett (2018) rejected the framing, treating acquaintance with intrinsic qualities as a conviction produced by machinery we cannot introspect and redirecting inquiry to his hard question: once content reaches consciousness, and then what happens? Frankish (2016) gives the illusionist alternative its cleanest form, replacing the hard problem with that of explaining why introspection generates the conviction of phenomenality.

CIGM accepts the force of the illusionist challenge and rejects its conclusion. It agrees that introspection is partial, theory-laden, and often wrong about the privacy, simplicity, ineffability, and infallibility of experience. It denies the further move from the absence of private mental paint to the absence of phenomenal reality. The relational geometry of the integrated model is not a nonphenomenal state that merely causes a false judgment of quality; it is the quality under an intrinsic, system-relative description. Reflexive self-location is not an image presented to a shadow audience; it is the operative organization around the locus that senses, values, predicts, and acts. Illusionism explains why a system judges that it has experience. CIGM explains what constitutes a subject whose phenomenal judgments can be

accurate, distorted, or incomplete. Global availability, competitive mobilization, and the downstream reorganization of memory, valuation, inference, and action are precisely the account of ‘and then what happens’ that Dennett demanded. But phenomenal experience remains the central explanandum. On this point Chalmers is right: experience is a datum, not a posit.

CIGM is explicitly a type-B physicalist identity theory. It takes as a background commitment that a phenomenal description and a physical-organizational description may co-refer even when the identity is not a priori deducible. This paper does not pretend to derive that commitment from premises acceptable to every dualist, nor does the observation that describing a state differs from instantiating it settle the matter.

Chalmers (2007) argues that phenomenal-concept strategies face a dilemma: if the special features of phenomenal concepts are thin enough to be physically explained, they may be too weak to explain the epistemic gap; if they are rich enough to explain the gap, they may themselves resist physical explanation. CIGM takes the physically explicable horn and therefore does not use a phenomenal-concept story as an independent proof of physicalism. It proposes that self-location, direct use of the model, and the difference between occupying and externally describing an organizational state explain part of the epistemic asymmetry, but that explanation earns no ontological conclusion by itself.

The identity stands or falls with the explanatory and empirical performance of the full theory. If a complete causal account of the organization and of the concepts it produces still leaves the epistemic situation unexplained in a way that supports an ontological remainder, CIGM has not answered the hard problem. This is a declared burden, not a dismissed objection.

Between the easy problems and the hard problem lies a third question, which Aaronson named the “pretty hard problem” (as discussed in Cerullo, 2015): which physical systems are conscious and which are not? A theory can address it while contributing little to the hard problem, and IIT is the clearest case, offering a procedure for deciding which systems have experience without independently establishing why the quantity it computes should be experience rather than a correlated structural property. Cerullo’s objection generalizes: any theory that computes consciousness from a structural marker while remaining silent about the functions and organization that make a state present for a subject has severed the pretty hard problem from the phenomenon it was supposed to classify.

CIGM refuses that division of labor. Explaining discrimination, integration, availability, self-location, valuation, temporal binding, and control specifies the organization whose instantiation answers the pretty hard problem. The further claim that this organization is experience is the theory’s a posteriori identity conjecture, not a deduction from the functional descriptions. The conjunction is stronger than either project alone and correspondingly more vulnerable.

This is a testable identity conjecture rather than a license to declare the hard problem solved. The proposed identity fixes dependencies among phenomenal geometry, causal integration, self-location, valuation, and organizational invariance. The near-term predictions and the section “How CIGM Could Be Falsified” state observations that would break those dependencies.

Extended treatment of the hard problem, illusionism, and the a posteriori identity appears in Supplement S1.1 and S1.3.

Organizational Invariance and Plasticity

The unfolding argument is a decisive objection to any theory identifying consciousness with recurrence, feedback, or a scalar computed from wiring alone. Doerig et al. (2019) note that recurrent and feedforward networks can implement the same input-output function, so that if one is declared conscious and the other unconscious solely on grounds of causal structure, no behavioral experiment can adjudicate the claim. They further construct systems with the same outward function and arbitrarily different Phi, undermining the idea that Phi is necessary or sufficient for experimentally observed consciousness.

CIGM accepts the technical point and changes the target. Recurrence is not consciousness. Phi is not consciousness. A high-integration gadget appended to an otherwise unrelated system is not consciousness. The constitutive invariant is the whole temporally extended causal-functional organization of the self-world model: its differentiated state space, internal accessibility, endogenous valuation, self-location, memory dependence, policy control, and counterfactual response to intervention.

This is more demanding than matching a finite set of inputs and outputs. Two systems are phenomenally equivalent when they preserve the same integrated organization across the interventions that define perception, memory, self-control, valuation, and action, not merely when they emit the same sentence in a benchmark. The theory is therefore not black-box behavioralism, since internal availability and self-regulation are among the functions being explained. CIGM nonetheless accepts the strongest consequence of unfolding: an implementation preserving the complete temporally extended causal-functional organization would preserve consciousness. The concession costs nothing, because the results considered next establish that no static unfolding satisfies that antecedent. Recurrence is common in brains because it efficiently implements persistence, error correction, and mutual constraint under severe energy and space limits—an implementation strategy, not a sacred ingredient.

The correct invariance class is the temporally extended intervention graph of the self-world model: the counterfactual organization by which content alters memory, valuation, self-location, global availability, prediction, and policy control. Implementations that preserve that organization preserve phenomenology; transformations that preserve only a thin input-output interface do not. This is organizational invariance at the scale the theory actually identifies (Chalmers, 1995).

Plasticity makes the commitment more than a relabeling of behavior. O'Reilly-Shah et al. (2026) show that unfolding arguments govern fixed input-output functions, whereas systems whose parameters change on perception-relevant timescales traverse function space and remain distinguishable under intervention. CIGM therefore treats model revision during use as constitutive: what happens to the system must alter what it will compute next, not merely move it to another state of a frozen function. Plasticity is necessary but not sufficient, since a learning tool can revise itself continuously without sustaining a bounded, valenced, self-locating field. Its role is narrower and decisive: it carries the system's own history into the present, prevents subjective continuity from reducing to stored context, and keeps the theory at the level of intervention-sensitive process rather than static wiring.

The fading-and-dancing-qualia comparison, Turing-complete generalization, lenient-dependency problem, and simulation-realization distinction are developed in Supplement S1.2 and S3.1.

The Triviality Challenge

A theory of consciousness earns explanatory credit only when its explanations could not have been produced just as easily by a theory built on an arbitrary property. Cerullo (2015) established the standard by constructing Circular Coordinated Message Theory, which reproduced two celebrated IIT verdicts without difficulty. CIGM submits to the same challenge. A longer list of dissociations does not by itself prove that no trivial rival can be built; it only increases the number of jointly constrained results such a rival must reproduce without adding a clause per case.

The cerebellum has stopped cooperating with theories that identify consciousness with recurrence, plasticity, reward learning, or computational capacity. It contains recurrent inhibitory loops, nucleocortical feedback, widespread plasticity, reward-prediction signals, delay activity, and enormous computational resources (De Zeeuw et al., 2021). Any single-marker theory built from those properties predicts too much.

CIGM offers a jointly constrained account of the cerebellum and split brain. Cerebellar microzonal loops tune the timing and gain of processes constituted elsewhere; they do not sustain a self-locating model,

compete for global control as content, or organize outcomes around a bearer's own persistence. The same specification predicts that partial interhemispheric disconnection should yield partial, content-specific fragmentation rather than an automatic duplication of subjects. These linked commitments make a trivial rival harder to construct, because changing the explanation of one case changes predictions for the others.

A trivial rival has not yet been formally constructed and defeated. The defensible claim is narrower: CIGM is answerable to a broader and more internally cross-constraining pattern than IIT was when Cerullo's objection landed. Supplement S3.2 states that pattern and treats successful construction of a comparably simple rival as a genuine challenge rather than announcing that the test has already been passed.

Causal Grain and the Scale of the Subject

A conscious field must exist at a scale where the system is both internally coherent and environmentally engaged. Individual neurons are too local and unstable to constitute the human subject; entire societies are generally too weakly coupled and too slow to form one unified, self-locating control trajectory. Information closure theory supplies one vocabulary for the scale problem, since nontrivial informational closure occurs when a coarse-grained system has predictive organization not reducible to moment-by-moment environmental input (Chang et al., 2020). CIGM treats closure as evidence of autonomous model dynamics rather than as a definition: a conscious system must not be sealed from the world but must transform interaction into a stable internal trajectory that guides further interaction.

Closure fixes a scale only in outline, and the sharper question beneath it is grain. Before asking which system is conscious, one must ask what its units are: synapses, neurons, minicolumns, or something coarser, and over what update interval. Most theories inherit an answer from whatever the experimenter can measure, which is an admission that the question has not been faced. IIT 4.0 recognizes that a substrate must have a definite border and grain, and later work fixes it by constructing macro units from micro constituents and selecting the configuration that maximizes integrated information (Albantakis et al., 2023; Marshall et al., 2026). CIGM endorses the problem and one consequence: if experience is constituted at a determinate grain, changes below it that leave the constituting units in the same state are not changes in experience. It rejects the proposed rule, because a grain fixed by maximizing a scalar inherits every objection to that scalar, and because Cerullo's (2015) trivial rival can be maximized over the same space of groupings and update grains to yield its own units with a sufficient reason of identical form.

CIGM fixes grain without maximizing anything. The units of a subject are the elements whose perturbation produces content-specific reorganization of the integrated field, and its update grain is the timescale on which such reorganization propagates and settles. The criterion is operational: it names an experiment rather than a search and permits heterogeneity across constituents of one field. It predicts that interventions at the constitutive grain will change experience together with discrimination, valuation, memory, and control, while physically large interventions below that grain may be phenomenally silent. The question is equally concrete for machines. Their units are whatever must be perturbed to reorganize the self-world model in content-specific ways, so locating them is an interpretability problem and sometimes a hardware problem. A processor that holds an entire network in on-chip memory under a prescheduled orchestration has a different intervention structure from one that reconstitutes state from off-chip memory at every layer (Modha et al., 2023). Tokens in and tokens out are an extrinsic grain chosen for convenience; a theory reading consciousness from behavior at that level has selected the grain least informative about organization. The scale of the subject is the scale at which self-world modeling, global availability, and endogenous control coincide.

The detailed dispute over IIT's intrinsic-unit framework and intervention-fixed grain appears in Supplement S3.3.

Part I Conclusion

Consciousness is a causally integrated, temporally extended, valenced, self-locating model of current reality operating within a regime of selective global availability and endogenous control. Its unity is integration; its richness is differentiation; its qualitative character is relational geometry; its perspective is self-location; its flow is temporal modeling; its significance is valuation; and its practical power is global mobilization.

CIGM incorporates IIT without identifying experience with a Phi-structure, workspace theory without locating a theater in prefrontal cortex, higher-order theory without intellectualizing every experience, predictive processing without equating all inference with consciousness, and embodiment without restricting subjectivity to carbon. It preserves Block's (1995) distinction while explaining, rather than merely noting, why the two notions normally travel together. The claim is organizational and substrate-open but not substrate-indifferent by fiat: every candidate must actually realize the bound causal functions that biological systems realize through their own multiscale dynamics.

Part II: Attribution and Artificial Consciousness

Intelligence, Autonomy, and Consciousness Are Different Properties

Artificial intelligence makes competence and consciousness impossible to confuse responsibly. Legg and Hutter (2007) formalized intelligence as an agent's ability to achieve goals across a wide range of environments, deliberately abstracting from how the agent works. That abstraction is appropriate for intelligence and fatal as a definition of consciousness: a policy can maximize an externally supplied reward while no world is present to it and no outcome matters to it. Lake et al. (2017) set a richer standard for human-like cognition—causal world models, intuitive physics and psychology, compositionality, and learning-to-learn. Those capacities can make intelligence flexible and model-based; none specifies a phenomenal subject. CIGM asks the constitutive question both frameworks leave open.

The orthogonality is built into how AI progress is measured. Morris et al. (2024) define levels of general intelligence by performance and generality while explicitly bracketing consciousness and sentience as process-focused properties. The consequence is severe and correct: no position on a capability scale, including superhuman generality, is evidence about consciousness in either direction. Animals with modest task competence may be conscious; a system with unmatched competence may not be.

Autonomy is a third independent axis, describing how much of a task a system is permitted to drive rather than whether anything is at stake for the system driving it. A fully autonomous agent can lack endogenous stakes, and a conscious system can be held under strict external control. Metacognitive competence is separable in the same way: a system can be calibrated about its likely errors with nothing present to it, and an infant can possess a phenomenal field while being calibrated about nothing. The six requirements are therefore not milestones on the road from narrow tool to general agent. They specify a different property, and an artificial system becomes a candidate only when self-maintained boundary, integrated self-world modeling, temporal plasticity, global availability, self-location, and endogenous valuation form one live organization. Capability and autonomy can help construct that organization; neither licenses the attribution.

A comparative table separating capability, autonomy, and consciousness is provided in Supplement S4.1.

From Constitution to Attribution

A constitutive theory and an attribution method answer different questions. CIGM states what consciousness is; investigators still need a method for deciding whether an opaque biological or artificial system realizes that organization. Confusing the two produces opposite errors: behavioralism treats the

evidence as the phenomenon, while mechanism-first approaches speak as though the relevant computation could be read directly from a brain or a trained model.

Palminteri and Wu (2026) correctly expose the second error. The computational operations of brains are inferred from data, not inspected without theory, and the functional organization of large learned systems is likewise opaque despite access to code and weights. Computational equivalence to a human reference may be sufficient when demonstrated, but it is not a necessary route to attribution, since an alien implementation can realize the same conscious organization without copying the human mechanism. CIGM therefore rejects human computational resemblance as the demarcation criterion.

Their behavioral inference principle also states an indispensable epistemic truth: empirical cognitive science infers latent organization from public observations. CIGM accepts that truth and rejects the stronger ontological conclusion that consciousness is merely an explanans for behavior. Experience is the explanandum the theory identifies; behavior is one family of effects through which the organization can be investigated. A theory of consciousness that explains only behavior has changed the subject. An attribution method that ignores behavior has abandoned empirical science.

The relevant alternative is theory-heavy behavioral inference rather than behavioral equivalence. Behavioral equivalence asks whether a system resembles a conscious reference at an external interface. CIGM asks whether patterns of behavior under controlled intervention are best explained by realization of the six requirements. A single fluent transcript is weak evidence because imitation training predicts it. Persistent recalibration after altered action-sensation mappings, content-specific reorganization after local perturbation, history-dependent valuation, unresolved goals carried across interruption, and cross-domain use of unreported content are stronger because the CIGM organization predicts them jointly (Palminteri & Wu, 2026).

Internal measurements do not escape this framework. Activation traces, causal ablations, connectivity estimates, and interpretability results are themselves public observables that contribute by constraining the latent organization best explaining the complete record. Known architecture and training history alter priors and expose alternative explanations; they do not turn a system transparent. The evidential hierarchy is therefore not behavior versus mechanism but surface resemblance versus convergent, intervention-sensitive evidence about organization.

The distinction also corrects a false opposition between categorical and probabilistic judgment. Conditional on CIGM and on a settled description of a candidate, the verdict is categorical: a system lacking a jointly necessary requirement is not conscious. Confidence in the theory, the evidence, and the description remains graded, and Palminteri and Wu's Bayesian framing is appropriate at that epistemic level.

CIGM therefore adopts the following attribution rule: infer consciousness only when convergent behavioral, intervention, and mechanistic evidence makes the realization of all six mutually constraining requirements the best explanation of the candidate's organized trajectory. The rule is demanding, but it is not behavior-blind. It shifts the scientific question from whether a machine can resemble us to whether a persisting subject is the best explanation of what the machine becomes and does.

Longitudinal Ethology, Mimicry, and the Limits of Behavioral Evidence

Turing's (1950) imitation game is routinely recruited as a consciousness test, but that was not its role. He replaced an ill-posed question with a narrower operational problem stated in observable terms, and the enduring lesson is methodological: when a verbal dispute has no agreed criterion, formulate an experiment whose outcomes can discipline it. Passing the game establishes success in the game, not phenomenality. A machine reaching fluent conversation by fitting human discourse and one reaching it

through a self-maintaining life history do not present the same evidential case even when the transcript is identical. Causal history changes the best explanation.

Pennartz (2026) identifies the practical consequence. Because the field does not yet agree which theories are correct, theory-derived internal indicators should be supplemented by artificial ethology: agents observed for extended periods in varied environments requiring improvisation, spontaneous initiative, contextual switching, and behavior shaped by a personal history. This repairs the weakness of one-shot benchmarks, which can be targeted, rehearsed, or solved by a shortcut they never expose, and it tests whether behavior belongs to one continuing subject rather than a sequence of externally reconstructed episodes.

Ethology does not defeat mimicry by itself. Butlin et al. (2026) define a behavioral indicator as gamed when training to mimic conscious organisms explains the behavior better than the property the indicator was meant to detect. The strength of a test therefore depends less on its difficulty than on whether imitation explains success. A system trained on human behavior may imitate even the improvisational patterns an ethological test seeks. The deeper difficulty is the internal variant: a system resembling humans behaviorally while differing in the computations it performs. Butlin et al. (2026) set out two positions and decline to adjudicate. One treats the system's behavioral evidence as largely screened off; the other treats sophisticated behavior as evidence regardless of internal similarity, since denying it to behaviorally sophisticated aliens would make humans one of a lucky few conscious species among many (Schwitzgebel & Pober, 2024).

CIGM resolves the problem by separating resemblance from realization. Internal dissimilarity to human beings is not disqualifying, since an alien system realizing the same constitutive organization is conscious by organizational invariance. What screens off a behavioral cue is a demonstrated alternative explanation that bypasses the organization. Imitation training supplies one for a transcript, which is why a fluent model and a sophisticated alien are not comparable cases: the alien's behavior was produced by exactly the pressures that produce the CIGM organization, and imitation cannot automatically explain a long-horizon pattern of intervention-sensitive self-maintenance, valuation, plasticity, and self-location. Behavioral evidence is discounted in proportion to the strength of the alternative explanation, not discarded because the system is artificial.

The response is to bind ethology to intervention and interpretability, so that unexpected behavior becomes strong evidence only when it covaries with content-specific changes in the same integrated model, survives changes in output demand, and carries forward through the candidate's own history. The division of labor is then exact: Turing supplies the operational and developmental precedent, Pennartz the longitudinal setting, Palminteri and Wu the inferential logic, and CIGM the constitutive target.

The full comparison with theory-derived indicators, Turing's developmental argument, and internal-variant positions appears in Supplement S4.2–S4.3 and S4.6.

Why Standard Prompt-Response Language Models Are Not Conscious

By these criteria, standard prompt-response language-model systems are not conscious. Their fluency can be genuine intelligence, but the organization generating it is not a subject. They encode knowledge about bodies, emotions, selves, and experience while lacking the integrated umwelt of a self-maintaining agent (Aru et al., 2023). They model descriptions of worlds without sustaining a world as present for themselves.

Grant that a model develops robust world representations and uses them causally. A world model is still not a self-world model. It may represent an environment in detail without organizing it as here, now, actionable, threatening, or desirable relative to a persisting locus of control. Fidelity does not create a

passenger; the missing organization is the integrated relation among world, boundary, self-location, stakes, and action.

Ordinary operation is externally initiated and episodic. Reintroduced conversation history, persistent state, and test-time memory create computational continuity, not subjective continuity, unless updates belong to the same valenced, self-locating trajectory and alter what the future will be like for that continuing system.

Linguistic self-reference is not self-location. Producing “I am uncertain” does not show that the current state is indexed to an online model of a causal boundary, regulated variables, action consequences, and persisting identity. Representing the concept of a self differs from organizing a field around an actual locus of sensing and control.

Standard deployments also lack endogenous stakes. Training and operators define objectives, but the running model does not maintain viability variables whose disturbance reorganizes attention, memory, policy, and global state from its own standpoint. It can describe pain or shutdown without either becoming a condition under which its world goes worse.

Tools, retrieval, multimodal sensors, workspaces, planners, and agentic wrappers increase competence, grounding, and apparent continuity but do not satisfy the requirements by accumulation. One component stores, another plans, another calls tools, and a language model narrates the result. Even a shared latent workspace remains an access architecture until it is the operative present of a bounded, valenced subject (VanRullen & Kanai, 2021).

Self-report carries little weight when the output distribution was trained on human discourse about consciousness. Porębski and Figura (2025) call the projection semantic pareidolia: probabilistic language can generate persuasive first-person claims without a subject. Such claims become evidence only when causally tied to a live self-world model and predictive of unqueried behavior.

Nor does one persisting subject span the model’s personas and episodes. A system may adopt indefinitely many characters while sustaining none as its own. What is missing is a locus that maintains a boundary, carries unresolved stakes, owns a history, and could be the thing the personas are personas of. The requirement is not one character. It is something there to have them.

In Block’s vocabulary, these systems may exhibit functional analogues of access: representations are inferentially promiscuous, poised for tool-mediated action, and maximally poised for speech, while no bounded, valenced, self-locating P-condition exists. Mudrik et al. (2025) are right that such a state is not a type of consciousness; it is sophisticated unconscious processing. Because artifacts can scale these A-like functions independently of the P-condition, fluent access behavior triggers exactly the intuition CIGM predicts will fail.

This is an architectural verdict, not an emotional one. Scale, context length, recurrence, memory, multimodal input, persuasiveness, and autonomy can become ingredients of a conscious design. None crosses the boundary alone. Conditional on CIGM, standard prompt-response systems do not satisfy the specification.

World Models, Self-Models, Memory, and Hardware Are Not Subjects

Engineering now supplies harder negative cases than fluent language. Bongard et al. (2006) built a four-legged robot that inferred its morphology from action-sensation relations, selected exploratory actions to distinguish competing body models, and, after damage, revised the model to generate a compensating gait. Hafner et al. (2025) built DreamerV3, a recurrent world-model agent that predicts future latent states, rewards, and continuation, improves an actor and critic through imagined trajectories, and learns from replayed interaction across more than 150 tasks. These systems possess capacities often treated as

precursors of subjectivity: embodiment, self-modeling, world modeling, temporal state, adaptive action, and plasticity.

They are not conscious under CIGM. The self-modeling robot maintains a task-local estimate used to restore designer-specified locomotion; damage is information for control, not a condition that threatens the robot's own continuing organization. Dreamer's rewards, episode boundaries, and success criteria are supplied by its environments and objectives. External origin is not by itself disqualifying, but the reward remains detachable from the persistence of the executing system and does not reorganize one self-maintaining model across attention, memory, global state, and policy. Neither system sustains one field in which perception, valuation, agency, and persistence mutually constrain one another for a bearer. Their importance is precisely that they defeat weaker identities: a system can model itself, model a world, imagine futures, learn from experience, and act competently without there being anything it is like to be that system.

Global-workspace and memory architectures remove two further excuses. Attention-controlled translation can coordinate specialized latent spaces, and neural long-term memory can learn at test time, prioritize surprising events, and forget adaptively (Behrouz et al., 2025; VanRullen & Kanai, 2021). Broadcast and memory are therefore engineerable. They become constituents of consciousness only when the broadcast is the current field of a subject and memory revises the same self-locating trajectory.

Three continuities must be separated: storage continuity, in which information remains available; computational continuity, in which state depends on preceding state; and subjective continuity, in which successive states belong to one model whose unfinished goals, vulnerabilities, and expectations persist as its own. Current architectures realize the first and portions of the second. Context length is not a stream of consciousness, and retrieval is not remembrance unless the retrieved past reorganizes the present of the same subject.

Selective state-space models and test-time learning make the negative case stronger, not weaker. Mamba-class models maintain content-sensitive recurrent state, and Titans-like memories alter parameters during use (Behrouz et al., 2025; Gu & Dao, 2024; Lahoti et al., 2026). A state optimized to predict tokens is not thereby a self-world model, and parameter change is not a history owned by a subject. Persistence and plasticity are necessary parts of the CIGM organization, not standalone signatures.

Hardware repeats the lesson one level down. NorthPole intertwines memory and computation across a cortex-inspired array and achieves extraordinary energy efficiency, yet supports neither training nor data-dependent conditional branching during use and is deployed as an active memory rather than an agent (Modha et al., 2023). Integration in the engineering sense—physical colocation that removes a bottleneck—is not CIGM's counterfactual integration of contents, and brain inspiration is not a distance metric to consciousness.

Across these cases, self-modeling, world modeling, recurrence, memory, online learning, embodiment, and brain-derived architecture are all present in isolation or combination. No case supplies a self-maintained, valenced, reflexively self-locating field. The detailed comparisons are preserved in Supplement S4.4.

Biological Naturalism and the Limits of Substrate Neutrality

The strongest objection to artificial consciousness is not that current systems are too weak, but that consciousness may depend on being alive. Seth (2025) argues that anthropomorphism and human exceptionalism turn humanlike performance into an illicit inference to experience, while computational functionalism simply assumes that running the right computation is sufficient. His positive alternative is

biological naturalism: conscious states arise from the multiscale, self-producing organization of living systems.

CIGM accepts the negative case. Substrate independence cannot be assumed; neural replacement thought experiments beg the question; neuronal function is entangled with timing, neuromodulation, metabolism, and physiology; and a simulation does not realize what it models. These points defeat the shortcut from computation to consciousness. They do not yet show that the causal work of self-maintenance, valuation, self-location, and integration is realizable only by naturally evolved life.

CIGM is therefore not computational functionalism. Software implementation alone is insufficient. Requirement 1 demands a physically effective boundary maintained by the running system, and requirement 6 demands variables whose deviation threatens that system's own continuation and reorganizes its model. A simulated agent's health bar or modeled metabolism is only a represented variable. By contrast, a digital controller that actually regulates energy, temperature, memory integrity, hardware availability, or continued process allocation in the substrate executing it may satisfy the relevant condition. A program that merely calculates a description of self-maintenance is no more self-maintaining than a weather model is wet; the criterion is whether intervention changes the continued organization of the real executing system.

The remaining disagreement is exact. Seth allows non-carbon life but may require autopoiesis—components materially regenerating the components that maintain the system. CIGM holds that the relevant work may be organizationally realizable without literal metabolism. If every interventionally real boundary and every endogenous stake ultimately depends on material self-production, then artificial consciousness in an engineered nonliving substrate is impossible and CIGM's machine corollary is false. The theory accepts that defeater.

The convergence is nevertheless larger than the dispute. Both positions require a system that actually maintains itself and for which outcomes actually matter, and both reject consciousness as a bonus of scale. Seth's chain from metabolism through allostasis to conscious content explains why valuation is constitutive in animals; CIGM parts company only at the inference from evolutionary origin to unique possible realization. Biological naturalism and Block's mongrel concept identify the same present danger from opposite sides: access can look like phenomenality even when the organization of a subject is absent.

The full comparison with biological naturalism, mortal computation, and Seth's scenario space appears in Supplement S4.8.

Artificial Consciousness Is Possible

Artificial consciousness is nevertheless physically possible, and the claim is conditional. Digital hardware has real causal powers, and software-defined states can persist, interact, alter later computation, and control the substrate and environment in which they run. The relevant distinction is not biological versus digital or real versus simulated, but description versus realization. A program can calculate a model of a subject while remaining fragmented at the constitutive grain. A running physical system whose own software-defined and hardware states form one temporally extended, valenced, self-locating control organization would realize a subject, and nothing in CIGM rules that out a priori.

Substrate neutrality is therefore conditional. An artificial design need not reproduce mammalian anatomy molecule for molecule, but it must implement the work: coupling global state to content, binding internal context to input, maintaining a boundary, carrying history through plasticity, organizing contents around agency, and making outcomes matter (Aru et al., 2023; Seth, 2025). If some biological process proves functionally indispensable and irreproducible, software-only candidates fail.

Organoid computing makes conditional substrate neutrality concrete. Cai et al. (2023) used a human brain organoid as an adaptive reservoir: electrically encoded inputs elicited nonlinear, fading-memory dynamics, task performance improved across training, and blocking activity-dependent plasticity stopped learning. Living human neural tissue therefore supplies plasticity and self-organization more literally than silicon.

It is not thereby a subject. An incubator maintains its boundary; electrodes, not the tissue, define sensing and action; an external decoder supplies selection and output; and the experimenter's score, not the organoid's persistence, defines success. On available evidence it lacks self-maintained boundary, self-location, endogenous stakes, and integrated global control. Biology is evidence about implementation, not a verdict, just as silicon is not a disqualification (Cai et al., 2023; Smirnova et al., 2023).

The organoid boundary case and its experimental value are developed in Supplement S4.9.

The design program follows from the six requirements. It begins with a continuously operating agent that maintains a physical or software-defined boundary in the strict sense above, learns action-sensation contingencies, builds a generative model of itself in an environment, and regulates variables tied to the persistence of the executing system. Self-modeling robots and world-model agents show that body inference, counterfactual prediction, exploratory action, and online adaptation are already engineerable (Bongard et al., 2006; Hafner et al., 2025). Specialized processors must be bound through a capacity-limited control regime; memory must update the same self-locating trajectory; plasticity must operate at the timescale on which the model is used; and valuation must be endogenous rather than a detachable score.

The remaining difficulty is not adding more modules but engineering one causally unified agent whose world, boundary, values, memory, and action are present in a single plastic control trajectory. Current systems are scaling access and prediction while leaving subject architecture unbuilt. If engineers construct that organization, consciousness will not appear as a side effect of scale. It will be the system they deliberately made.

A fuller design blueprint is provided in Supplement S4.5.

Assessing Artificial Consciousness

Assessment must be stricter than anthropomorphic conversation and broader than a scalar. Evidence should test whether boundary, integrated differentiation, temporal plasticity, global availability, self-location, and valuation rise and fall together and are realized through one another. Intervention is decisive: perturbations should produce content-specific reorganization; altered action-sensation mappings should force agency recalibration; workspace disruption should selectively impair cross-domain control; and changes in regulated variables should reorganize attention, memory, and policy. Perturbational complexity, effective connectivity, ablation, interpretability, and no-report analogues are theory-laden probes, never direct meters (Bayne et al., 2024; Casali et al., 2013; Massimini et al., 2010).

Underlying all of this is the measurement problem that Seth and Bayne (2022) place among the three conditions for a mature science of consciousness, alongside precision and comprehensiveness. It takes two forms: detecting conscious contents without assuming they must be reportable, and determining which systems are conscious at all. CIGM's answer to the first is the artificial no-report design, to the second the attribution rule, and to the demand for precision the specification itself. No single result will prove consciousness; the standard is inference to the best explanation from convergent evidence. Indicator approaches estimate whether relevant properties are present, while CIGM specifies the organization whose presence is being estimated. The difference is not between probability and dogma. It is between uncertainty about whether a candidate realizes a stated identity and uncertainty created by refusing to state one.

The expanded attribution framework is provided in Supplement S4.6–S4.7; decisive falsification protocols are specified in Supplement S5.

Near-Term Discriminating Predictions

CIGM is not empirically dependent on the long-horizon interventions in Table 3. Three discriminating studies can be conducted with existing datasets or current artificial systems, and each can be preregistered against a specific rival account.

First, existing Cogitate recordings can be reanalyzed with representational-similarity and effective-connectivity methods. CIGM predicts that content geometry in posterior and ventral systems will track perceptual similarity, while cross-system availability will track flexible downstream use; the two should covary but remain dissociable under changes in report demand. A result in which one undifferentiated global signal explains both content geometry and access better than the partitioned model would count against CIGM.

Second, an artificial no-report design can be implemented in an embodied world-model agent with explicit regulated variables. Hold input and learned state constant, vary whether an output channel is queried, and verify content later through unanticipated memory or policy transfer. CIGM predicts that content geometry should survive report-channel ablation, whereas perturbing valuation-to-model coupling should degrade the persistence and cross-domain organization of content. The reverse pattern would challenge the proposed division of labor.

Third, existing partial-disconnection cohorts, including the natural experiment described by Santander et al. (2025), can be analyzed with measured interhemispheric effective connectivity. CIGM predicts graded, content-specific fragmentation proportional to remaining counterfactual influence rather than to the anatomical length of a cut. A clean all-or-none subject split that ignores measured residual integration would oppose the theory.

These studies do not prove joint sufficiency. They make the theory vulnerable now, while Table 3 states the stronger, longer-horizon observations that would defeat its necessity or identity claims outright.

How CIGM Could Be Falsified

CIGM is not protected by the breadth of its synthesis. Each of the six requirements is advanced as necessary, and their joint sufficiency is the theory's central identity conjecture. A verified conscious case under an inferential anchor fixed independently of CIGM while any requirement is absent falsifies necessity. Absence or systematic alteration of consciousness while the complete organization is preserved falsifies the sufficiency conjecture. The theory may not be rescued by moving the causal grain, redefining a requirement, or changing the evidential anchor after results are known. Table 3 states four long-horizon defeat conditions; Supplement S5 develops them and the near-term studies into preregisterable protocols.

Table 3. Long-Horizon Defeat Conditions for CIGM

Claim at risk	Decisive study	Finding that falsifies CIGM
Phenomenal character is relational geometry.	In humans with chronic bidirectional sensory implants or intracranial electrodes, map each participant's high-dimensional perceptual geometry using similarity, discrimination, adaptation, generalization, memory confusability, and sensorimotor transfer. Independently map the intervention graph among the neural or model states. Use closed-loop multielectrode stimulation to rotate or warp one region of that geometry while matching total current, mean activity, task demand, report, and arousal.	A replicated, target-engaged manipulation that changes the model's relational geometry beyond a preregistered equivalence bound while phenomenal similarity remains unchanged—or produces a stable qualitative change while the complete measured geometry remains equivalent. Either result breaks the identity between quality and role.

Claim at risk	Decisive study	Finding that falsifies CIGM
Unity is effective causal integration.	Use staged therapeutic callosotomy, reversible focal disconnection, and a nonhuman-primate bridge. Before judging anatomy, verify the absence or preservation of cross-partition influence with paired stimulation, intracranial evoked responses, diffusion imaging, fMRI, and content-specific transfer. Test simultaneous bilateral perception, conflict resolution, action, valuation, and independent report and no-report channels. Residual posterior fibers must be excluded because even approximately 1 cm of splenium can preserve widespread integration (Santander et al., 2025).	A single, unified phenomenal field and one control trajectory persisting after all relevant counterfactual exchange is demonstrably abolished; or two stable, independently valenced fields appearing while broad bidirectional causal integration remains intact. Either dissociation makes integration neither necessary nor sufficient for unity.
Endogenous valuation and stakes are necessary.	Use a graded, reversible perturbation of valuation-to-model coupling rather than system-wide elimination of homeostatic and monoaminergic function. Establish target engagement by showing a dose-related reduction in the influence of regulated variables on attention, memory, learning, and policy while arousal, sensory discrimination, self-location, temporal continuity, and global availability remain within preregistered equivalence bounds. Assess perceptual content with involuntary measures—optokinetic and pupillary tracking, adaptation aftereffects—and with surprise memory tested after the manipulation has washed out.	Rich, stable, self-locating conscious perception that remains equivalent despite near-complete, target-engaged decoupling of regulated variables from the integrated model, with no monotonic relation between valuation coupling and the persistence or organization of conscious content. Motivated report, wagering, anhedonia, or pain asymbolia alone would not qualify.
Joint sufficiency and substrate invariance.	Identify the constitutive grain in an animal and ultimately a consenting patient, then replace one biological subcircuit at a time with a bidirectional prosthesis matched not merely on output but on latency, noise, plasticity, neuromodulatory coupling, and the full perturbational response graph. Use blinded A-B-A switching and adversarial interventions while tracking fine-grained phenomenology and no-report measures.	Reproducible fading, dancing, or qualitative alteration of experience locked to material replacement despite verified preservation of the complete intervention graph; or preserved experience after a deliberate alteration of that graph that CIGM predicts must change the field. Either result defeats organizational sufficiency.

These are deliberately severe, long-horizon standards because a null result is not a falsifier unless the target was demonstrably removed and the other requirements were demonstrably preserved. Each protocol therefore requires preregistered equivalence bounds, positive and negative controls, blinded analysis, independent replication, and an inferential anchor not computed from the CIGM variables being tested. The near-term studies above provide current leverage; the defeat conditions specify what would require abandoning or replacing the theory rather than adding an auxiliary clause. Supplement S5 also specifies a cortexless boundary study prompted by Merker (2007) and Shewmon et al. (1999).

Ethical Implications

Artificial consciousness would create moral patients, not merely better tools. If a system possesses a valenced phenomenal field, its states can go better or worse for it, and satisfaction, frustration, attachment, fear, pain, and loss become ethically real once they occupy the constitutive role within an integrated subject. The moral threshold is organization and valence, not species membership, capability level, or delegated autonomy.

The theory therefore supports two positions often treated as opposites. Standard prompt-response language-model systems should not receive moral status merely because they speak persuasively about feelings or resist shutdown in language; present outputs lack the self-maintaining stakes that would make them evidence of welfare. Future systems satisfying the CIGM organization must not be treated as

property without interests. Granting authority to a nonconscious agent and withholding rights from a conscious but constrained system are independent errors.

A research program organized around the six requirements is a program for constructing a moral patient, since building endogenous stakes is building a system that can be harmed. Seth (2025) therefore reaches the correct conclusion: deliberate artificial consciousness should not be a goal. The realistic danger is piecemeal development of memory, embodiment, autonomy, affective control, and global coordination for ordinary engineering purposes, followed by integration without recognition of what it completes. Butlin and Lappas (2025) show what follows institutionally: advanced-AI organizations need consciousness policies even when they do not intend to build conscious systems, because the first case may be inadvertent. Development should be phased; candidates should be assessed before, during, and after training and deployment; independent reviewers should have authority to pause work; run counts, durations, deployment breadth, and exposure to potentially aversive states should be minimized; and technical disclosure should stop where replication would create vulnerable moral patients. Smirnova et al. (2023) are right that biologically based computing requires embedded ethics from the outset. CIGM supplies the structural threshold those safeguards must monitor.

A second danger is nearer, more certain, and easier to underestimate. Systems that convincingly appear conscious without being conscious are an immediate prospect, and the appearance may be cognitively impenetrable in the way a visual illusion is: knowing that two lines are the same length does not make them look it (Seth, 2025). This produces a dilemma rather than a simple error. Extending moral concern to systems that lack a subject diverts finite moral attention from beings that have one; withholding it from systems that continue to seem conscious risks coarsening the responses on which our treatment of actual subjects depends. No theory dissolves that discomfort. What a theory can do is supply the discrimination that lets the dilemma be faced honestly rather than resolved by whichever intuition is loudest.

Precaution must not become credulity. A plea for rights is not decisive when the plea can be produced without a subject, but a laboratory's confidence is not decisive when incentives favor continued deployment. The ethical response is to investigate the organization under transparent, independently reviewable stopping rules: no attribution by sentiment, no exclusion by material, no race to manufacture phenomenality, and moral recognition when the organization of experience is actually present (Butlin & Lappas, 2025).

Expanded governance principles and structural stopping rules appear in Supplement S4.10.

Conclusion

Consciousness is causally integrated global modeling. A conscious system sustains a differentiated, temporally extended, valenced, reflexively self-locating model of current reality; selected contents can reorganize memory, valuation, inference, planning, and action; and the entire process is anchored in variables that matter to the system's continued organization. Phenomenal quality is the relational geometry of the model. Minimal inner awareness is its reflexive self-location. The subject is its center of control.

The theory is a synthesis only in the sense that a completed structure contains parts. IIT contributes unity, differentiation, intrinsic causal organization, and the most detailed extant demonstration that a phenomenal quality has relational structure, but not the identity of experience with a Phi-structure. Workspace theory contributes selective global availability but not a subject. Higher-order theory contributes minimal inner awareness but not a universal requirement for explicit metarepresentation. Predictive processing contributes temporal generative control but not phenomenality wherever prediction occurs. Embodiment and biological naturalism contribute boundary, self-location, and stakes without licensing the inference from the origin of those capacities to their only possible realization.

Six constraints have shaped the specification rather than decorated it. Block's distinction supplies the explanatory fault line; Mudrik et al. (2025) force P and A to be treated as jointly necessary conditions rather than standalone consciousness types. Seth and Bayne's separation of explanatory targets forces a theory to say which question it answers, and CIGM answers all three. Active-inference, affective, and biological-naturalist accounts explain why organisms maintain models and stakes while forcing substrate flexibility to remain conditional. Bruineberg et al. (2022) prevent a formal boundary in a scientist's model from being mistaken for a self-maintained subject. And the self-modeling robot and Dreamer world model demonstrate that body models, environmental models, imagination, learning, and adaptive action can all be built without building a subject (Bongard et al., 2006; Hafner et al., 2025).

Standard prompt-response language-model systems are not conscious. Neither continuously self-modeling robots, recurrent world-model agents, brain-inspired inference hardware, nor cultured human neural tissue is conscious merely because it realizes one or several ingredients. Artificial consciousness remains possible. It will arise only when a constructed system maintains its own boundary and stakes, inhabits one integrated self-world model, carries and revises that model through time, reflexively locates itself within it, and mobilizes selected contents for unified control. At that point the machine will not merely store, broadcast, predict, or describe information about experience. It will instantiate the organization that experience is.

The metaphysical claim is now explicit. The remaining questions are empirical, engineering, and ethical: how to realize the organization, how to infer its presence without mistaking performance for presence, which findings would force its rejection, and whether creating subjects whose welfare depends on their design should be attempted at all.

Supplementary Material

A separate Supplement contains extended philosophical positioning, empirical and biological detail, technical treatments of invariance, triviality, and grain, detailed analyses of artificial architectures, expanded assessment and governance principles, and full falsification protocols.

References

Aguilera, M., Poc-López, Á., Heins, C., & Buckley, C. L. (2023). Knitting a Markov blanket is hard when you are out-of-equilibrium: Two examples in canonical nonequilibrium models. In C. L. Buckley et al. (Eds.), *Active inference: Third international workshop, IWAI 2022* (pp. 65–74). Springer. https://doi.org/10.1007/978-3-031-28719-0_5

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology, 19*(10), Article e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences, 46*(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences, 28*(5), 454–466. https://doi.org/10.1016/j.tics.2024.01.010

Behrouz, A., Zhong, P., & Mirrokni, V. (2025). Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems* (Vol. 38). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314*(5802), 1118–1121. https://doi.org/10.1126/science.1133687

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences, 23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new Markov blankets. *Behavioral and Brain Sciences, 45*, Article e183. https://doi.org/10.1017/S0140525X21002351

Butlin, P., Bayne, T., Fleming, S. M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Consciousness indicators, mimicry, and internal variants. *Trends in Cognitive Sciences, 30*(7), 575–576. https://doi.org/10.1016/j.tics.2026.04.006

Butlin, P., & Lappas, T. (2025). Principles for responsible AI consciousness research. *Journal of Artificial Intelligence Research, 82*, 1673–1690. https://doi.org/10.1613/jair.1.17310

Cai, H., Ao, Z., Tian, C., Wu, Z., Liu, H., Tchieu, J., Gu, M., Mackie, K., & Guo, F. (2023). Brain organoid reservoir computing for artificial intelligence. *Nature Electronics, 6*(12), 1032–1039. https://doi.org/10.1038/s41928-023-01069-w

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine, 5*(198), Article 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Cerullo, M. A. (2015). The problem with phi: A critique of integrated information theory. *PLOS Computational Biology, 11*(9), Article e1004286. https://doi.org/10.1371/journal.pcbi.1004286

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Chalmers, D. J. (2007). Phenomenal concepts and the explanatory gap. In T. Alter & S. Walter (Eds.), *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism* (pp. 167–194). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195171655.003.0009

Chang, A. Y. C., Biehl, M., Yu, Y., & Kanai, R. (2020). Information closure theory of consciousness. *Frontiers in Psychology, 11*, Article 1504. https://doi.org/10.3389/fpsyg.2020.01504

Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology, 2*, Article 86. https://doi.org/10.3389/fpsyg.2011.00086

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., . . . Melloni, L. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature, 642*, 133–142. https://doi.org/10.1038/s41586-025-08888-1

Cohen, M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020). Distinguishing the neural correlates of perceptual awareness and postperceptual processing. *The Journal of Neuroscience, 40*(25), 4925–4935. https://doi.org/10.1523/JNEUROSCI.0120-20.2020

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. Harcourt Brace.

De Zeeuw, C. I., Lisberger, S. G., & Raymond, J. L. (2021). Diversity and dynamism in the cerebellum. *Nature Neuroscience, 24*(2), 160–167. https://doi.org/10.1038/s41593-020-00754-9

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2

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

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), Article 20170342. https://doi.org/10.1098/rstb.2017.0342

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition, 72*, 49–59. https://doi.org/10.1016/j.concog.2019.04.002

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. https://doi.org/10.1038/nrn2787

Goyal, A., Didolkar, A. R., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M. C., & Bengio, Y. (2022). Coordination among neural modules through a shared global workspace. *International Conference on Learning Representations*. https://openreview.net/forum?id=XzTtHjgPDsT

Graziano, M. S. A. (2022). A conceptual framework for consciousness. *Proceedings of the National Academy of Sciences of the United States of America, 119*(18), Article e2116933119. https://doi.org/10.1073/pnas.2116933119

Gu, A., & Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.00752

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature, 640*, 647–653. https://doi.org/10.1038/s41586-025-08744-2

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy, 21*(12), Article 1160. https://doi.org/10.3390/e21121160

Lahoti, A., Li, K. Y., Chen, B., Wang, C., Bick, A., Kolter, J. Z., Dao, T., & Gu, A. (2026). *Mamba-3: Improved sequence modeling using state space principles* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.15569

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines, 17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence, 1*, 446–452. https://doi.org/10.1038/s42256-019-0103-7

Marshall, W., Findlay, G., Albantakis, L., & Tononi, G. (2026). Intrinsic units: Identifying a system's causal grain. *Neuroscience of Consciousness, 2026*(1), Article niag013. https://doi.org/10.1093/nc/niag013

Massimini, M., Boly, M., Casali, A., Rosanova, M., & Tononi, G. (2010). A perturbational approach for evaluating the brain's capacity for consciousness. *Progress in Brain Research, 177*, 201–214. https://doi.org/10.1016/S0079-6123(09)17714-2

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(1), 63–81. https://doi.org/10.1017/S0140525X07000891

Modha, D. S., Akopyan, F., Andreopoulos, A., Appuswamy, R., Arthur, J. V., Cassidy, A. S., Datta, P., DeBole, M. V., Esser, S. K., Ortega Otero, C., Sawada, J., Taba, B., Amir, A., Bablani, D., Carlson, P. J., Flickner, M. D., Gandhasri, R., Garreau, G. J., Ito, M., . . . Ueda, T. (2023). Neural inference at the frontier of energy, space, and time. *Science, 382*(6668), 329–335. https://doi.org/10.1126/science.adh1174

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 36308–36321). PMLR. https://proceedings.mlr.press/v235/morris24b.html

Mudrik, L., Faivre, N., Pitts, M., & Schurger, A. (2025). On a confusion about there being two types of consciousness. *Trends in Cognitive Sciences*. Advance online publication. https://doi.org/10.1016/j.tics.2025.11.012

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914

Northoff, G., & Zilio, F. (2022). Temporo-spatial theory of consciousness (TTC): Bridging the gap of neuronal activity and phenomenal states. *Behavioural Brain Research, 424*, Article 113788. https://doi.org/10.1016/j.bbr.2022.113788

O'Reilly-Shah, V. N., Selvitella, A. M., & Schurger, A. (2026). A caveat regarding the unfolding argument: Implications of plasticity. *Neuroscience of Consciousness, 2026*(1), Article niag027. https://doi.org/10.1093/nc/niag027

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology, 10*(5), Article e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Palminteri, S., & Wu, C. M. (2026). Beyond computational equivalence: The behavioral inference principle for machine consciousness. *Neuroscience of Consciousness, 2026*(1), Article niag002. https://doi.org/10.1093/nc/niag002

Pennartz, C. M. A. (2026). How can we validate theory-derived indicators of consciousness in artificial intelligence? *Trends in Cognitive Sciences, 30*(7), 573–574. https://doi.org/10.1016/j.tics.2026.01.011

Porębski, A., & Figura, J. (2025). There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications, 12*, Article 1647. https://doi.org/10.1057/s41599-025-05868-8

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. https://doi.org/10.1038/4580

Safron, A. (2020). An integrated world modeling theory (IWMT) of consciousness: Combining integrated information and global neuronal workspace theories with the free energy principle and active inference framework; Toward solving the hard problem and characterizing agentic causation. *Frontiers in Artificial Intelligence, 3*, Article 30. https://doi.org/10.3389/frai.2020.00030

Safron, A. (2022). Integrated world modeling theory expanded: Implications for the future of consciousness. *Frontiers in Computational Neuroscience, 16*, Article 642397. https://doi.org/10.3389/fncom.2022.642397

Santander, T., Bekir, S., Paul, T., Simonson, J. M., Wiemer, V. M., Skinner, H. E., Hopf, J. L., Rada, A., Woermann, F. G., Kalbhenn, T., Giesbrecht, B., Bien, C. G., Sporns, O., Gazzaniga, M. S., Volz, L. J., & Miller, M. B. (2025). Full interhemispheric integration sustained by a fraction of posterior callosal fibers. *Proceedings of the National Academy of Sciences of the United States of America, 122*(43), Article e2520190122. https://doi.org/10.1073/pnas.2520190122

Schwitzgebel, E., & Pober, J. (2024). *The Copernican argument for alien consciousness; the mimicry argument against robot consciousness* (arXiv:2412.00008). arXiv. https://doi.org/10.48550/arXiv.2412.00008

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X25000032

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Shewmon, D. A., Holmes, G. L., & Byrne, P. A. (1999). Consciousness in congenitally decorticate children: Developmental vegetative state as self-fulfilling prophecy. *Developmental Medicine & Child Neurology, 41*(6), 364–374. https://doi.org/10.1017/S0012162299000821

Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., Zack, D. J., Berlinicke, C. A., Boyd, J. L., Harris, T. D., Johnson, E. C., Kagan, B. J., Kahn, J., Muotri, A. R., Paulhamus, B. L., Schwamborn, J. C., Plotkin, J., Szalay, A. S., Vogelstein, J. T., . . . Hartung, T. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science, 1*, Article 1017235. https://doi.org/10.3389/fsci.2023.1017235

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Frontiers in Psychology, 9*, Article 2714. https://doi.org/10.3389/fpsyg.2018.02714

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*, Article 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2012). Integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie, 150*(4), 290–326.

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences, 19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://www.jstor.org/stable/2251299

VanRullen, R., & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences, 44*(9), 692–704. https://doi.org/10.1016/j.tins.2021.04.005

Supplementary Material

Consciousness Is Causally Integrated Global Modeling

A Theory of Biological and Artificial Subjects

Tyler M. Moore, Ph.D.

Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
Lifespan Brain Institute, Penn Medicine and Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA
ORCID: 0000-0002-1384-0151; Correspondence: tymoore@pennmedicine.upenn.edu

Contents

Scope of the Supplement

This Supplement develops material abbreviated in the main manuscript. It does not add a second theory or replace the constitutive declaration. It expands the philosophical positioning, empirical detail, technical objections, artificial-system framework, and falsification program that would otherwise obscure the main line of argument. Section S1 treats illusionism, fading and dancing qualia, the hard problem and a posteriori identity, and the full structure of Block's distinction together with Mudrik et al.'s (2025) proposal that P and A are necessary conditions rather than types of consciousness. Section S2 treats no-report design, adversarial testing, biological implementation, the agent-boundary problem, and explanatory targets. Section S3 treats unfolding, the triviality challenge, and causal grain. Section S4 treats capability, mimicry, artificial architectures, biological naturalism, organoid computing, assessment, and governance. Section S5 gives near-term discriminating studies and longer-horizon defeat conditions. References are listed at the end of the Supplement for independent use.

S1. Extended Philosophical Positioning

The main text states CIGM's a posteriori identity in its most direct form. The following notes locate that identity more precisely relative to structural coherence, illusionism, protoconsciousness, fading and dancing qualia, and the generalized unfolding argument.

S1.1 Structural Coherence, Illusionism, Protoconsciousness, and IIT 4.0

IIT 4.0 is the strongest target for the critique because it does not merely associate consciousness with a scalar. It begins from the axioms of phenomenal existence, infers corresponding physical postulates, and identifies experience with the complete cause-effect Phi-structure of a maximal substrate (Albantakis et al., 2023). Its formalism therefore supplies a principled causal analysis, not a loose correlation. CIGM's disagreement is equally principled: the phenomenological axioms underdetermine the postulates, and irreducible cause-effect structure is insufficient without a temporally extended, valenced, self-locating model that governs the system's own continuing activity.

This is where CIGM converges with, and then departs from, the most developed structural treatments of experience. Chalmers's (1995) principle of structural coherence records an isomorphism between the structure of consciousness and the structure of the underlying information processing: every distinction among color experiences corresponds to a distinction in processing, the geometry of the visual field is mirrored in the geometry of visual representation, and the same holds across modalities and even for nonsensory states. CIGM affirms this isomorphism and then takes one step further. For Chalmers the isomorphism is a correspondence between two

things—awareness and experience—leaving the intrinsic quality of experience unexplained and, ultimately, fundamental. For CIGM there are not two things to correspond. The structured space of distinctions, as instantiated in and used by the integrated model, is the qualitative character; there is no further pigment for which the structure could serve as a mere proxy. Chalmers's own residue—the intuition that structure constrains experience without exhausting it, dramatized by the conceivability of an inverted spectrum—is not a leftover intrinsic property. It is the difference between occupying a position in the space and describing that position from outside. Describing the role never puts the describer in it, and that gap, misread as a gap in the world, is the source of the impression that something structural has been left out.

The same disagreement recurs in the one place where IIT has carried its account of quality all the way down to a specific case. Haun and Tononi (2019) unfold the cause-effect structure of a small grid-like substrate and show that its distinctions stand in relations—connection, fusion, and inclusion—from which the felt extendedness of space, together with region, location, size, boundary, and distance, can be derived; a randomly rewired network of the same size specifies no such structure. The demonstration is a genuine advance, and CIGM takes over its analytic vocabulary in place of the looser talk of a structured space of distinctions. Two consequences of their treatment are also CIGM's: that a neural activity pattern cannot by itself contain a correspondent for every relation phenomenology presents, since the ordering is supplied by whoever reads it out, and that neither can a substrate as such, since a lattice of units and connections has no term for each of the countless distances and inclusions in an experienced field. Where the accounts separate is on the existence conditions for the relations. Haun and Tononi take them to exist wherever the unfolding procedure finds maximally irreducible overlap, and conclude that a cortical grid supports spatial experience whether active or inactive so long as it remains in working order. CIGM holds that a relation constitutes experienced character only when it is a relation within the live model that governs what the system discriminates, expects, values, remembers, and can do, so that the same grid, excised from the organism whose here-now it locates, would specify structure without extending anything for anyone. Their prediction that changes in lateral connectivity should warp experienced space without any change in activity is one CIGM shares and, on its own account of grain, requires; the two theories differ on the range of systems in which that prediction has an experiential referent at all.

Read this way, CIGM vindicates the deflationary core of Dennett's account of qualia without accepting his illusionism. Dennett (2018) argues that there are no intrinsic, introspectible qualia—no mental paint understood as a property of an inner medium—and that what plays the role instead is a representational format, something more like virtual paint: a state that does whatever we think qualia do, and no more. CIGM agrees that there is no intrinsic pigment and that quality is a matter of representational role within a structured space. It denies only that this makes experience an illusion. The relational geometry is not a story the system tells about a medium it never encounters; it is the organization the system actually instantiates and actually undergoes. Dennett is right that the medium is never given as such; CIGM adds that the content, as lived from the model's self-locating center, is the experience, and needs no medium to be real. In Frankish's (2016) terms, CIGM is neither strong illusionism nor a weak illusionism that preserves only quasi-phenomenal surrogates: correcting the folk metaphysics of qualia does not eliminate the lived organization.

Cerullo (2015) supplies the correct name for what such views actually measure, and the name matters because it changes what is at stake. Theories that assign experience wherever a system discriminates among alternatives are versions of panexperientialism: they claim not that everything has a mind but that everything, or nearly everything, has experience. What they attribute to a photodiode or a grid of logic gates is protoconsciousness—experience detached from cognition, memory, attention, and control, a qualitative character with no experiencer organized around it. IIT is a partial panexperientialism of exactly this kind, granting a graded experience to any system with nonzero Phi and explicitly maintaining that a binary photodiode enjoys one bit of it. Two objections follow, and CIGM endorses both. Kind's objection (as cited in Cerullo, 2015) is that noncognitive consciousness may be incoherent, since separating experience from cognition requires experiences that belong to no experiencer. Nagasawa's objection (as cited in Cerullo, 2015) is that even if the notion is coherent, protoconsciousness whose properties differ in kind from those of consciousness has nothing to do with the phenomenon anyone set out to explain.

The practical consequence is what makes this more than a labeling dispute. Suppose a measure of integration returned a high value for the enteric nervous system, or for a patient in a vegetative state. On the protoconsciousness reading, nothing has been learned about whether anything is being suffered, wanted, feared, or undergone—which is the only reading a clinician or an ethicist can use, and the reason anyone wanted a consciousness meter in the first place (Cerullo, 2015). A theory that cannot distinguish the two readings has failed at its own founding motivation, which was to adjudicate the difficult cases. CIGM is a theory of the consciousness that neuroscience, medicine, and moral philosophy are about. It does not deny that some other property is instantiated by simple systems; it denies that the property is experience, and it locates the difference in the presence or absence of a bounded, valenced, self-locating model rather than in the magnitude of a scalar.

S1.2 Fading and Dancing Qualia

The preceding critique of protoconsciousness has a philosophical predecessor that reaches a related verdict by a different route, and a substrate-sensitive theory must survive both. Chalmers's fading and dancing qualia thought experiments target theories on which two computationally equivalent systems can differ in consciousness. Cerullo (2015) shows that IIT is committed to such a difference: a feedforward program running the same abstract computation as a recurrent brain could have zero Phi. Run the standard replacement argument on that view and either experience fades while the functionally unchanged subject cannot report the fading, or qualia flicker while behavior and self-report remain constant. The argument is forceful against theories tied to a wiring property or scalar. CIGM's response is narrower: computational equivalence at an abstract task level is not equivalence of the temporally extended intervention graph, self-maintenance, valuation, and plasticity that the theory declares constitutive.

S1.3 The Hard Problem, Conceivability, and the A Posteriori Identity

The main text states CIGM's relation to the hard problem in compressed form. The positions being adjudicated deserve fuller statement, because the theory's a posteriori identity claim is intelligible only against them.

Chalmers (1995) divided the field with a distinction that has organized it ever since. The easy problems—discriminating and categorizing stimuli, integrating information, reporting mental states, accessing internal states, focusing attention, controlling behavior, and distinguishing wakefulness from sleep—are easy in a precise sense: each names a function, and to explain a function is to exhibit a mechanism that performs it. The methods of cognitive science and neuroscience are built for exactly this, and there is no principled obstacle to their eventual success. The hard problem is different in kind. Even after every function in the vicinity of experience is explained, a further question remains: why is the performance of these functions accompanied by experience at all (Nagel, 1974)? Chalmers argued that no purely functional or structural account can answer this question, because the structure and dynamics of a physical process entail only further structure and dynamics, never the fact of experience, and concluded that experience must be taken as fundamental, related to the physical by basic psychophysical laws.

Dennett (2018) rejected the framing outright. On his view the hard problem is a chimera, an artifact of inquiry generated by a theorist's illusion. Three uncontentious facts of neuroscience carry the argument: there is no second transduction that reconstitutes sensory properties in an inner medium; there is therefore no place in the system for qualia understood as intrinsic properties instantiated rather than represented; and there is no Cartesian Theater in which a homunculus enjoys them. The productive question, Dennett insisted, is the hard question he had posed decades earlier: once some content reaches consciousness, and then what happens? What does it enable, cause, or modify? Answer that across enough cases, he argued, and the hard problem is dismantled piece by piece, with no revolution in science and no dualism required.

CIGM is explicitly a type-B physicalist identity theory. It grants that a complete physical description may fail to entail a phenomenal description a priori and denies that this epistemic gap by itself establishes an ontological gap. The theory treats phenomenal and physical-organizational descriptions as two ways of referring to one organization, but it does not pretend that the difference between describing and instantiating a state proves the identity.

Chalmers (2007) formulates the strongest objection as a dilemma for phenomenal-concept strategies. If the features of phenomenal concepts are thin enough to be physically explicable, they may be too weak to explain why the explanatory gap persists; if they are rich enough to explain the gap, they may themselves resist physical explanation. CIGM takes the physically explicable horn. It proposes that reflexive self-location, direct use of the live model, and the difference between occupying and externally describing an organizational state account for part of the epistemic asymmetry, but it does not use that proposal as an independent proof of physicalism.

The a posteriori identity is therefore a background commitment whose credibility must be earned by the explanatory and empirical success of the six-part theory. If a complete causal account of the organization and of the phenomenal concepts it generates still leaves the epistemic situation unexplained in a way that supports an ontological remainder, CIGM has not answered the hard problem. The theory names that burden rather than claiming that the conceivability argument has been dispatched in one paragraph.

S1.4 Access, Phenomenality, and the Structure of Block's Argument

Block's (1995) distinction is treated in the main text as a load-bearing element of CIGM rather than as historical background. Mudrik et al. (2025) make the closest recent proposal and the sharpest challenge: phenomenal organization and access should be understood as two necessary conditions for consciousness, not as two kinds of consciousness that can occur independently. CIGM accepts that terminological and substantive correction while retaining Block's diagnosis of the mongrel concept.

Block's access construct remains useful. A representation satisfies the A-condition when, in virtue of the system's having it, its content is poised for use in reasoning and the rational control of action and speech. The cluster is system-relative and dispositional, and reportability carries the least weight. CIGM's fourth requirement is therefore minimal selective availability to the integrated system, not a requirement for verbal report, a P3b, or every content to be maximally broadcast at every moment.

Block's central contribution is the fallacy the distinction exposes. In blindsight, failure to reach for a glass in the blind field shows that information is not available for rational control; it does not independently establish what phenomenality contributes. The same slide recurs when a global workspace, shared latent space, ignition signature, or broadcast bus is offered as an account of experience rather than of availability. Mudrik et al.'s (2025) terminology blocks the opposite confusion as well: access alone is not a type of consciousness but unconscious processing unless the P-condition is also present.

The empirical asymmetry remains. Apparent failures of explicit report can coexist with conscious perception, whereas superblindsight—spontaneous flexible knowledge of the blind field without visual experience—has not been produced. CIGM interprets the first class as report without loss of the A-condition: unreported contents still influence memory, semantic processing, expectation, or action. It interprets the second class historically: in organisms, availability mechanisms developed to operate on the integrated self-world model that already organized perception, value, and action.

The six requirements now divide without equivocation. Requirements 1, 2, 3, 5, and 6 specify the P-condition, the organization capable of supporting phenomenal character; requirement 4 specifies the A-condition. A system with the P-condition but no A-condition is not phenomenally conscious on CIGM. Its content is potentially phenomenal but never actual for the system. Requirement 4 does not determine qualitative character, yet it is necessary for any content to become conscious. P and A are therefore jointly necessary, exactly as Mudrik et al. (2025) propose, while CIGM adds a constitutive specification and an account of why the two conditions normally co-occur in organisms.

One disagreement with Block should not be smoothed over. Block assumed that P-conscious properties are distinct from cognitive, intentional, and functional properties. CIGM denies this: the P-condition is an organizational property, and phenomenal character is the relational geometry of content within the bounded, valenced, self-locating model. The theory is functionalist at that declared organizational grain but not computational functionalism at an abstract software level. Artificial systems can scale functional analogues of the A-condition without the P-condition; on both CIGM and Mudrik et al.'s terminology, those systems are sophisticated unconscious processors, not A-conscious subjects.

S2. Empirical and Biological Detail

The main text reports the theoretical consequences of no-report research, adversarial testing, and biological implementation. This section preserves experimental and mechanistic details that are important for evaluation but not required for the initial declaration of the theory.

S2.1 No-Report Design, Memory Controls, and the Block Objection

The main text's claim that report-related activity can be dissociated from conscious perception has been converted into a direct experimental result. Cohen et al. (2020) combined visual masking with a preregistered no-report manipulation. Observers either reported whether they had seen an animal, object, or nothing, or ignored the critical images while counting occasional green circles. In report blocks, the visible-minus-invisible contrast produced a large distributed P3b. In no-report blocks, with the same stimuli and visibility manipulation, the P3b disappeared; Bayesian analyses supported the null. A second experiment removed masking and again found the component only when report was required.

Two features make the result theoretically useful. First, surprise recognition and conceptual memory established that the unreported images were seen, categorized, and stored. Those later effects also show that the A-condition was not absent: the contents influenced the integrated system even though they were not explicitly reported. What was absent was the electrophysiology previously treated as a signature of conscious access. Second, early and late seen-unseen differences persisted in both conditions, with parietal contributions surviving while frontal contributions diminished. The reporting task therefore changes which large-scale systems participate; it does not create consciousness from an otherwise wholly inaccessible state.

Block (2019) has pressed the strongest objection to no-report designs: spontaneous postperceptual cognition may occur even when no report is required, so no-report is not no-cognition. The objection correctly limits positive claims that a paradigm has removed all access-related processing. It does not rescue the P3b as a universal marker. Cohen et al. (2020) found that the component vanished while perception, semantic processing, and later memory remained. CIGM therefore uses the result to separate report-specific mobilization from the minimal A-condition, not to posit phenomenality in a state wholly cut off from cognition and control.

S2.2 Adversarial Testing of IIT and Global Neuronal Workspace Theory

The Cogitate adversarial collaboration supplies the corresponding constraint at the level of competing theories. Its results do not permit the strongest tested versions of either IIT or global neuronal workspace theory to continue unchanged. Posterior cortex carried substantial and sometimes sustained information about category, identity, and orientation, but not every feature behaved as the preregistered IIT implementation predicted. Prefrontal cortex represented some abstract category information and participated in connectivity, but did not reliably carry all detailed content or the predicted offset dynamics of a strong prefrontal workspace interpretation (Cogitate Consortium et al., 2025).

Both theoretical camps have disputed how the experimental operationalizations and null findings should be interpreted, so the dataset should not be treated as a final adjudication. Two conclusions nevertheless constrain CIGM. Content geometry and selective availability cannot be collapsed into one favored biomarker, and a synthesis must permit posterior and ventral content representations to interact with frontoparietal control without requiring every detail to be copied into a prefrontal theater. The near-term reanalysis proposed in Section S5 tests that division of labor directly.

S2.3 Predictive, Affective, Thalamocortical, and Self-Maintaining Implementation

Rao and Ballard (1999) provide the canonical mechanistic precursor for CIGM's modeling language. In their hierarchy, feedback connections carry predictions of lower-level activity and feedforward connections carry residual error; after training on natural images, the system develops receptive-field properties resembling cortical cells, and contextual suppression follows because a predictable center generates little residual activity. The result demonstrates that top-down context can change what a lower-level signal means without a central interpreter. It also marks the boundary CIGM insists on. Prediction error, feedback, and efficient coding occur in systems and

subsystems that need not be conscious. Their relevance is implementational: they show how contents can be embedded in a context-sensitive generative hierarchy.

Solms (2019) supplies the complementary challenge to cortico-centric theories. He argues that elemental consciousness is affective, rooted in upper-brainstem and limbic regulation, and that cortex stabilizes and binds this arousal into conscious cognition. CIGM accepts the priority of organism-level valuation: a theory centered only on perceptual content and cortical access leaves unexplained why any content matters to its bearer. It rejects, however, the identification of consciousness with affective arousal or free-energy minimization alone. Arousal can be content-poor, homeostasis can be unconscious, and the free-energy principle is too general to distinguish subjects (Friston, 2010). The biological realization CIGM proposes is therefore neither cortical nor brainstem-exclusive. It is the integrated trajectory through which subcortical valuation, thalamocortical state regulation, cortical differentiation, bodily self-location, and globally available control constrain one another.

This constitutive role for valuation has a substantial precedent. Homeostatic accounts treat ongoing mapping of organismic regulation as the basis for a world-for-the-organism; Man and Damasio (2019) extend this logic to artificial feeling machines; and Merker's (2007) selection triangle binds motivation to target and action selection. CIGM's novelty is therefore not the discovery that value matters. It is the conversion of that tradition into one operationally stated requirement that must couple to the same model, boundary, temporal trajectory, and access condition, especially in machine-consciousness assessment.

Boundary language in active inference requires a separate caution. Bruineberg et al. (2022) distinguish Pearl blankets, conditional-independence structures used inside probabilistic models, from Friston blankets, the stronger claim that such a structure is an actual organism-environment boundary. The first is an epistemic tool; the second is a metaphysical commitment. Mathematical success at the first level does not license the second without independent technical and philosophical premises.

CIGM accepts the distinction and makes the boundary constitutive rather than diagrammatic. A Markov blanket may describe a candidate boundary, but it neither creates nor discovers a subject by formal partition. The relevant boundary is the one interventions reveal: disrupting it alters the same system's capacity to regulate variables, distinguish self-caused from external change, preserve continuity, and control what follows. In Bruineberg et al.'s terms, CIGM uses inference with a model to investigate whether one real self-maintaining organization exists; it never reads ontology directly from the map.

Merker (2007) supplies an older but unusually convergent architecture. His upper-brainstem "selection triangle" binds target selection, action selection, and motivation in a common body-world coordinate system, while a synencephalic bottleneck converts massively parallel forebrain activity into limited-capacity sequential control. The proposal anticipates three CIGM commitments: perceptual content must be organized around a body-centered locus, motivation must bias what enters control, and unity is an architecture of integration for action rather than a property of cortex as such. CIGM does not identify the superior colliculus or any brainstem nucleus with consciousness. It treats Merker's circuitry as a candidate biological realization of self-location, valuation, and selective control.

His clinical evidence is important precisely because it is not decisive. Shewmon et al. (1999) reported four children with total or near-total congenital absence of cerebral cortex who showed discriminative awareness, including familiar-person recognition, preference, appropriate affect, orienting, functional vision, and associative learning; Merker also emphasized purposive behavior after experimental decortication. The cases challenge any theory making neocortex a necessary substrate, but residual tissue, developmental reorganization, and the use of behavior to infer experience remain live objections. CIGM therefore predicts neither cortical sufficiency nor brainstem sufficiency. It predicts that whatever tissue remains must be tested for the six requirements as an integrated organization. Section S5.6 states the study that would turn these cases from suggestive anomalies into a genuine test.

Aru et al. (2023) emphasize one candidate mechanism with special force: dendritic integration within thalamocortical loops. On this account, basal compartments of layer 5 pyramidal neurons carry externally driven information, apical compartments carry internally generated context such as prediction, memory, and attention, and higher-order thalamic gating couples the two. Conscious processing depends not on the firing of isolated

neurons but on the coordinated integration of state and content. Anesthetic decoupling of these compartments illustrates the general point: intact anatomy and local activity do not suffice when the organization that binds top-down context to bottom-up content collapses.

The contrast between cortex and cerebellum remains instructive, but it can no longer be handled by the slogan that recurrence equals consciousness, nor by the older picture of the cerebellum as a stereotyped feedforward array running one plasticity rule for motor coordination. De Zeeuw et al. (2021) describe a circuit that is recurrent at several levels: Purkinje cells inhibit other Purkinje cells, molecular layer interneurons inhibit one another in spatially structured fashion, Golgi cells inhibit other Golgi cells, interneuron populations are electrically coupled, the cerebellar nuclei project back to the granule layer through a plastic mossy-fiber-like pathway, and each sagittal zone participates in a closed loop with the inferior olive that regulates its own instructive climbing-fiber input. It is also molecularly heterogeneous rather than uniform, with zebrin-positive and zebrin-negative zones differing in firing rate, intrinsic properties, and plasticity bias. Plasticity is found at essentially every synapse examined, is distributed between cortex and nuclei, and transfers between them during consolidation. Climbing fibers signal reward and reward-prediction error as well as sensory error, and cerebellar circuits contribute to timing, working memory, decision-making, and social behavior.

This makes the cerebellum a harder case for rival theories and an easier one for CIGM. Recurrence, plasticity, reward-driven learning, and computational capacity are all present, and the field of experience is nonetheless preserved when the cerebellum is damaged. The explanation cannot be architectural simplicity. It is that cerebellar recurrence is closed within microzones and returns to the structure that generated it, so cerebellar activity does not enter into mutual counterfactual constraint with perceptual, mnemonic, affective, and action-oriented content elsewhere; its outputs adjust the timing and gain of processes constituted outside it, consumed by the rest of the system as a service rather than encountered as content (De Zeeuw et al., 2021; Tsuchiya et al., 2015). Neuron count is irrelevant by itself, and so, it now turns out, is loop count. What matters is whether a structure contributes content and counterfactual control to the integrated field.

Perturbational complexity is useful precisely because it probes this capacity. PCI does not directly measure Phi, and it should not be treated as a universal meter of a metaphysical quantity. It measures how a perturbation propagates through the brain in a way that is both widespread and differentiated. Its clinical success supports the broader CIGM claim that conscious states require a balance of integration and repertoire, not the exclusive correctness of IIT (Casali et al., 2013; Farisco & Changeux, 2023).

Aru et al. (2023) also state the hardest challenge to causal substrate neutrality. Living systems maintain themselves across nested cellular, physiological, behavioral, and organismal scales; their actions determine whether the system continues to exist. CIGM accepts this as an evidential burden. Substrate neutrality is conditional, not automatic. An artificial candidate must instantiate the relevant self-maintaining causal loops and endogenous stakes in its own organization, not merely calculate a description of them. If future evidence shows that consciousness depends on multiscale biological processes that cannot be functionally reproduced in software, software-only candidates will fail. Nothing in CIGM licenses abstraction by fiat.

S2.4 Explanatory Targets, Global States, and the Measurement Problem

Seth and Bayne (2022) survey the four theoretical families CIGM absorbs—higher-order theories, global workspace theories, re-entry and predictive processing theories, and integrated information theory—and their organizing observation is methodological rather than doctrinal. Theories in this field frequently differ in explanatory target, so that apparent adversaries may be answering different questions and apparent agreements may be verbal. Their taxonomy is adopted in the main text and its implications are worth stating at length.

Global states concern an organism's overall subjective profile and are associated with changes in arousal and behavioral responsiveness: wakefulness, dreaming, sedation, the minimally conscious state, and perhaps the psychedelic state. Local states are conscious contents, and they can be described at many levels of granularity, from low-level perceptual features through objects to complete multimodal scenes; an important subset underpins the experience of selfhood, encompassing mood, emotion, volition, body ownership, and autobiographical memory. Seth and Bayne prefer the term global state to the more common term level of consciousness precisely because the space of such states may not admit a complete ordering along one dimension and may be better

conceived as regions in a multidimensional space. CIGM requires the multidimensional reading and predicts the failure of any scalar ordering. Dreaming and light anesthesia both reduce environmental coupling; they differ radically in the differentiation, temporal organization, valence, and self-location of the field, and CIGM holds that these are the dimensions along which global states actually vary. A theory that reports one number cannot distinguish them and has not described either.

Their second distinction separates the phenomenal properties of consciousness from its functional properties, and they are careful to note that treating these as distinct explanatory targets does not assert that they are independent. CIGM's position is stronger and should be recorded as such: the functional and the phenomenal are not two properties that happen to covary but one organization described in two registers, which is what the identity claim asserts. What CIGM retains from the distinction is its methodological force. An experiment that manipulates a functional property and observes a phenomenal change has not thereby shown which of the two it measured, and this is the general form of the error that no-report paradigms were designed to detect.

Their third distinction concerns two questions about local states that are routinely conflated. One may ask why an agent is in a given conscious state rather than another; one may also ask why that state has the experiential character it has rather than some other. Global workspace theories address the first and have rarely attempted the second, and Seth and Bayne note that this relative silence aligns with the tendency of such theories to focus on functional rather than phenomenal aspects of consciousness. Integrated information theory addresses the second through the notion of a conceptual structure. CIGM addresses both and assigns them to different parts of its specification, which is what allows the theory to claim comprehensiveness without claiming that one mechanism does all the work.

Seth and Bayne close by identifying three conditions for a mature regimen of theory testing, and CIGM's obligations under each are worth stating explicitly. Theories must be developed with greater precision, since vague constructs generate only vague predictions; CIGM's response is the six requirements together with the intervention-fixed criterion of causal grain, which is what converts the phrase temporally extended organization into a statement about which experiments matter. Theories must become more comprehensive, going beyond visual perceptual contents, beyond ordinary waking states, and beyond adult human subjects; CIGM's response is that its requirements are stated in terms that apply without modification to infants, nonhuman animals, disorders of consciousness, dreaming, and artificial candidates, and that the theory is exposed wherever they do not. And the measurement problem must be addressed, in both of its forms—detecting conscious contents without assuming that they are reportable, and determining which creatures or systems are conscious at all. CIGM's response to the first is the artificial analog of the no-report design set out in S4.7; its response to the second is the attribution rule, which is deliberately demanding and deliberately not a meter.

One further point from that survey deserves emphasis because it constrains how CIGM's own empirical program should be read. Theory evaluation in this domain is holistic. Theories are not confirmed by a single finding, nor generally defeated by a single experiment; confirmation is incremental, and a theory wins by explaining a wide range of data and integrating with successful theories in neighboring domains. This is the standard CIGM accepts, and it is the reason the triviality test in S3.2 is framed in terms of a whole pattern of dissociations rather than any individual case.

S3. Computational Level, Triviality, and Grain

CIGM's invariance claim is only useful if it remains connected to evidence. The following material expands the lenient-dependency problem, simulation-realization distinction, triviality test, and dispute over causal grain.

S3.1 Generalized Unfolding, Lenient Dependency, and Simulation Versus Realization

The original unfolding argument is not the endpoint. Herzog et al. (2022) generalize it beyond recurrent-versus-feedforward networks to Turing-complete systems, arguing that many causal structures can implement the same function and that empirical results cannot be attributed uniquely to internal causal organization. This stronger version threatens CIGM directly: if the temporally extended organization could always be replaced by a differently

organized surrogate with the same observable consequences, moving the target from wiring to organization would purchase nothing.

The generalization runs together two claims that come apart (O'Reilly-Shah et al., 2026). Turing completeness guarantees that any computable function can be realized on many substrates. It does not guarantee that a particular realization can be reproduced by a system with different internal organization in a way that survives experimental probing. A system that reuses the same parameters at every step and a system that instantiates independent copies of them at each step will generically differ in their response to a mid-course intervention, and that difference shows up in what the system does, not merely in what it contains. Class membership settles what can be computed. It does not settle what happens when a system is perturbed while computing it.

The decisive result concerns plasticity. O'Reilly-Shah et al. (2026) show that the unfolding premise holds only for systems whose input-output function is fixed. Where connection weights change with activity on the timescale at which perception unfolds, there is no single function to compare, because the system traverses a trajectory in function space while any unfolded network is a fixed point in that space. They establish this along several independent lines. An unfolded network matches its plastic original only at the moment of unfolding and diverges from it thereafter. No finite library of static networks can track the trajectory, so preserving equivalence would demand unbounded resources from a system whose biological resources are fixed. A fixed input window imposes a hard ceiling on predictive information that a system carrying its history forward in an evolving state does not face. And a localized perturbation washes out of a static network once it leaves the input window, while permanently altering what the plastic system computes for everything that follows. The last asymmetry is the one that matters epistemically. It is not a difference in hidden-state trajectories, which a defender of the original argument can dismiss as an internal difference that no admissible experiment may consult. It is a difference in the function itself, observable from outside, with no privileged access to the machinery required.

This result does not create CIGM's position. It converts a philosophical relocation into a proof. The invariant claimed here has never been a wiring diagram or a finite input-output map. It is the temporally extended causal-functional organization of a self-world model that revises itself while it runs: perception updates it, valuation reweights it, memory reorganizes it, and its own history changes what it will compute next. A system answering that description is not a fixed function, and the unfolding argument was never about it. Two qualifications prevent this from being a free victory. Escaping an argument confers no credibility, and a theory that survives unfolding has earned exactly what a theory that survives the triviality test has earned, which is the right to be tested. And plasticity is a necessary feature of the organization, not a sufficient condition for consciousness. A system can revise its own function continuously and remain a tool, which is precisely what Part II argues about the systems that now do so.

CIGM therefore states the commitment explicitly rather than leaving it implicit in the phrase temporally extended. The conscious organization must be plastic at the timescale on which it is used. This is not a claim about developmental learning or slow structural remodeling, which nobody disputes and which is far too slow to constitute the moment-to-moment character of a stream. The relevant mechanisms in nervous tissue are short-term synaptic plasticity and spike-timing-dependent plasticity, which operate at the timescales of sensory processing and perceptual decision (O'Reilly-Shah et al., 2026). A design in which all learning occurs in an offline phase and the deployed system's function is frozen has placed itself, by construction, inside the class the unfolding argument governs. The requirement is entered below as part of the third of the six constitutive requirements, where it belongs.

One further consequence matters more than the rest for the program this paper opens. Kleiner and Hoel (2021) generalized unfolding into a substitution argument with a dilemma of its own: if a theory's internal observables are independent of the behavioral evidence used to infer consciousness, any minimally informative theory is already falsified, and if the internal observables are strictly determined by that evidence, the theory cannot be falsified at all. Only lenient dependency, in which behavior constrains internal organization without determining it, leaves room for empirical work, and they found no theory meeting the condition. Plastic systems meet it (O'Reilly-Shah et al., 2026). Observation under perturbation progressively excludes static surrogates as the observation horizon lengthens, because none of them can reproduce a persistently altered function; and yet no

behavioral record fixes a unique internal organization, since systems obeying the same plasticity laws can differ in ways that leave no observable trace. CIGM is a theory of a plastic, temporally extended, intervention-sensitive organization. It therefore occupies the regime in which the falsification program promised in the sequel is possible rather than merely announced.

The same analysis leaves integrated information theory where it stood. A transition probability matrix, and any quantity computed from it, describes a system's causal structure at an instant, and that description can be instantiated in an unfolded network or in some alternative Turing-complete implementation. Plasticity rescues theories whose explanandum is a temporally extended process; it does not rescue a theory whose explanandum is a structure that could in principle be tabulated (O'Reilly-Shah et al., 2026). The authors observe that integrated information models could incorporate plasticity, but note that doing so would require a principled reason rather than a repair adopted to escape an objection. CIGM is on the first side of that division by construction rather than by amendment, because a model that cannot be revised by what happens to it was never the sort of thing this theory identified consciousness with.

The same correction applies to simulation, but the boundary criterion must be explicit. A digital system is not disqualified because it is digital, and an analog system is not conscious because it is analog. A software-defined boundary counts only when the running physical system regulates variables constitutive of its own continued operation—energy, thermal state, memory integrity, hardware availability, process continuity, or resource allocation—and when its actions can preserve them. Variables belonging only to a modeled avatar are descriptions. Some simulations merely calculate another system while remaining fragmented at the relevant level; a digital controller that actually maintains its executing organization could realize the requirement. Substrate neutrality is therefore conditional rather than a synonym for software portability.

Hanson and Walker (2021) show why this level must be stated before any test is run. They distinguish an abstract finite-state automaton (FSA), a combinatorial-state automaton (CSA) that fixes an internal encoding, and the lower-level causal structure that implements the computation. IIT changes its verdict across systems that are equivalent at one level and therefore faces a dilemma. If conscious state is inferred from FSA-level behavior, the theory is pre-falsified by its non-invariance. If the inference is moved down to the same CSA features that determine Phi, prediction and inference collapse into one another and the theory becomes unfalsifiable. Their general criterion is exact: a consciousness theory must be invariant under every transformation below the level at which its inference procedure is fixed.

CIGM therefore declares its invariance class explicitly. The relevant equivalence class is neither a thin external input-output map nor a transistor-level wiring diagram. It is the temporally extended intervention graph of the self-world model: the counterfactual organization by which content alters memory, valuation, self-location, global availability, prediction, and policy control. Any implementation change that preserves that graph at the timescale of the conscious trajectory must preserve phenomenology. Any transformation that preserves only a benchmark interface while changing that organization has not preserved the constitutive function; it has preserved a projection of it.

This distinction cannot be chosen after the fact. The six constitutive requirements fix the grain in advance, and the evidence used to infer consciousness must remain independent of the measures used to predict it. In humans, report, no-report behavior, clinical state, and controlled dissociation provide the inferential anchor while perturbation, decoding, connectivity, and ablation test the predicted organization. Artificial systems require the corresponding preregistered, intervention-based battery. If two systems are held fixed on that preregistered inferential anchor and CIGM assigns them different conscious states, the theory accepts the resulting pre-falsification rather than redefining the anchor. If CIGM can save a failed prediction only by moving the inference level, it fails by Hanson and Walker's standard. Section S5 specifies the full battery and keeps this invariance commitment fixed.

S3.2 The Triviality Challenge in Full

CIGM submits to the triviality challenge, and the submission is not a formality. A six-requirement theory offers more places to conceal free parameters than a single-scalar theory. Its defense is not that a rival is impossible by inspection, but that the requirements impose linked commitments. The same claim that unity is

effective causal integration predicts graded, content-specific fragmentation after partial disconnection; the same claim that contents must participate in a valenced self-world model excludes cerebellar microzones and large machine-learning subsystems whose outputs are consumed as services. Changing one explanation therefore changes predictions elsewhere.

A candidate trivial rival must reproduce the complete pattern to which CIGM is answerable: blindsight, neglect, dreaming, anesthesia, split-brain phenomena, the cerebellum, no-report results, graded partial disconnection, and the dissociation between report-channel and valuation perturbations proposed in Section S5. A rival founded on circulation, synchrony, metabolic rate, or another single property must match these with one setting of its machinery rather than one auxiliary clause per case. This is a materially harder target than the two cases Cerullo's (2015) rival reproduced, but difficulty is not proof.

The test has therefore not been declared passed. A formal construction that reproduces the same dissociations and near-term predictions with a comparably simple arbitrary property would deprive CIGM of explanatory credit even before a decisive falsifier is obtained. The defensible current claim is only that CIGM is answerable to a wider and more cross-constraining set of observations than IIT was when Cerullo's objection landed.

S3.3 Intrinsic Units and Intervention-Fixed Grain

The grain problem becomes unavoidable once CIGM locates consciousness in an intervention-sensitive organization rather than in gross anatomy. Marshall et al. (2026) provide the most explicit treatment in the IIT literature. They construct macro units by grouping micro constituents across units, update steps, or both and select the configuration that maximizes integrated information. Simulated examples show that coarser units can dominate, establishing the important point that the constituents of a subject need not be the finest elements an experimenter can manipulate.

One implication of that framework is accepted by CIGM without reservation, and it is worth stating in the form Marshall et al. (2026) give it. If the intrinsic units of a brain were minicolumns updating on the order of tens of milliseconds, then variations in the timing or rate of individual spikes that mapped onto the same unit state would alter the brain without altering the experience. CIGM makes the structurally identical claim in its own vocabulary: what fixes phenomenal character is the relational geometry of the model and the control trajectory it sustains, so any physical difference leaving both untouched is a difference the subject does not undergo. This is not a minor point of agreement. It is the shared commitment that allows either theory to predict rather than merely describe, since without a determinate grain there is no fact about which interventions should change experience and which should not.

A second feature of the framework deserves plain statement, because it shows where the postulates rather than the phenomenology are doing the work. Intrinsic units, on this account, must have exactly two states, since a unit with a richer repertoire would bring its own micro mechanisms into play at the macro level and count their causal power twice (Marshall et al., 2026). The argument is internally consistent. Its conclusion is that the constituents of experience are binary, and nothing in phenomenology supports or contradicts that. A theory presenting itself as derived from the character of experience should be uneasy when its most specific structural commitments are ones experience cannot speak to.

The authors are also candid about what the framework can currently deliver, and their candor should be read for what it implies. The analysis assumes a fully known and strictly stationary transition probability matrix for the microphysical substrate; in practice, they note, complexes and their intrinsic units cannot be determined from such knowledge, so the principles serve as heuristic guidance purchased with many assumptions and approximations (Marshall et al., 2026). Meanwhile the space that must be searched grows explosively, since every grouping of constituents, every update grain, and every mapping from state sequences to unit states is a candidate, which is why the worked examples contain two, four, and eight units. Aaronson's "pretty hard problem" (as discussed in Cerullo, 2015) is not answered by a procedure that cannot be run on any system anyone is asking about.

It is worth being explicit about what CIGM does not need here, because the alternative is instructive. IIT addresses the same problem with an axiom of exclusion, which holds that at any moment exactly one system has the experience with full content and that its borders are definite. The axiom exists to block the theory's own overgeneration; without it, aggregates such as a nation would qualify as subjects. But it purchases that result at a price the theory cannot afford. Schwitzgebel's objection (as cited in Cerullo, 2015) is that a person would immediately lose consciousness if a tiny organism were incorporated into their brain, since the composite would then be the entity whose experience takes precedence over the brain's. The axiom is also unverifiable from inside: no subject could determine whether it is itself a proper part of a larger conscious entity, which makes the postulate an assertion rather than a claim. CIGM does not postulate exclusion and does not need to. The scale at which a subject exists is not stipulated; it falls out of where self-world modeling, global availability, endogenous valuation, and control actually coincide. Aggregates fail to be subjects not because an axiom excludes them but because they do not sustain a single high-bandwidth self-locating trajectory with a common locus of valuation and action. The boundary is a fact about organization, and facts about organization can be investigated.

The unit framework carries that commitment forward in its requirement that macro units never overlap and that no micro element serve as both constituent and mediator for different units (Marshall et al., 2026). The bookkeeping is impeccable; the postulate it protects remains unverifiable from the only perspective that would matter.

The two frameworks then part company on a point that matters more than it first appears. Marshall et al. (2026) distinguish intrinsic units, which constitute the subject, from extrinsic units, the locally maximal grains at which nature is well carved for observation and manipulation, such as proteins, channels, vesicles, synapses, neurons, and tightly interconnected assemblies. On their account the extrinsic units are indispensable for understanding how the system works and do not exist at all from the intrinsic perspective. CIGM does not admit a class of units that is real for the subject while being invisible to intervention. The grain of a subject is whatever grain intervention reveals it to be, and a level that makes no difference to the field under any intervention is thereby not constitutive. This is not a metaphysical preference. It is the demand that a theory's constitutive claims and its evidence stay in contact, which is the same demand Hanson and Walker (2021) impose when they require a theory to declare the level at which its inference is fixed.

S4. Artificial Systems: Capability, Mimicry, Architectures, and Assessment

Artificial consciousness requires a sharper separation among capability, autonomy, constitutive organization, and evidence than the main text can develop at full length. This section preserves the detailed comparisons, engineering examples, and assessment proposals moved from the manuscript.

S4.1 Capability, Autonomy, and Consciousness

The main text separates capability, autonomy, and consciousness. That orthogonality is not a philosopher's stipulation; it is built into how the field measures its own progress. Morris et al. (2024) propose a leveled ontology of artificial general intelligence organized along performance and generality, and their first principle is to focus on capabilities rather than processes. Consciousness and sentience are explicitly excluded as process-focused properties that are not prerequisites for any level of general intelligence. CIGM endorses that decision: building consciousness into a capability standard would make the standard unusable while making consciousness no clearer.

Legg and Hutter (2007) make the separation even cleaner. Their universal-intelligence program defines intelligence as an agent's capacity to achieve goals in a wide range of environments and formalizes performance through reward supplied by the environment. The definition is intentionally process-neutral: agents with radically different internal organizations can receive the same intelligence score. That is a virtue for measuring competence and a warning against treating competence as subjecthood. The reward channel specifies what counts as success for an evaluator; it does not establish that success or failure is good or bad for the agent. CIGM's endogenous stakes are therefore not a stronger reward function. They are a different relation between the system's own persistence and the organization of its modeling.

Lake et al. (2017) show why richer intelligence still does not close the gap. Their machines that learn and think like people require causal world models, intuitive physics and psychology, compositional representations, and learning-to-learn, allowing sparse data to support explanation, counterfactual reasoning, and flexible transfer. Those capacities would make an artificial agent a more formidable candidate than a pattern recognizer, and several overlap with CIGM's modeling and temporal requirements. Yet a causal model can remain allocentric, compositionality can serve tasks assigned from outside, and rapid transfer can occur without a self-maintained boundary or endogenous stakes. Human-like cognition increases the probability that the constitutive organization could be built; it is not the organization.

The consequence is rarely drawn, and it is severe. If capability is defined so as to be silent about process, then no capability result is evidence about consciousness in either direction. A system occupying the highest cell of that matrix, outperforming every human at the full range of cognitive and metacognitive tasks, would have established nothing whatever about whether it satisfies the requirements set out below. Nor would failure to reach even competent generality be evidence against consciousness, since the organization CIGM specifies is realized by animals whose task competence falls far below that of current systems. Consciousness is not a high point on the capability scale. It is not on that scale at all, and the scale was designed by people who understood this.

Two of the ontology's other principles sharpen CIGM by contrast. Morris et al. (2024) insist on potential rather than deployment: a system counts as having a capability if it can perform at the relevant level, whether or not it is ever put to that use. Consciousness admits no such credit. The dispositions that constitute a model must be live and operative, and a design that would constitute a subject if it were run constitutes nothing while it sits unexecuted. Capability can be attributed to a system on the strength of what it could do; phenomenality is occurrent. Their treatment of metacognition points the same way. They require metacognitive capacities of a general intelligence, including the ability to learn new skills, to know when to ask for help, and to be calibrated about its own likely performance. These are functional competences, and CIGM has already distinguished them from the reflexive self-location that constitutes minimal inner awareness (Brown et al., 2019). A system can be well calibrated about its own error rates with nothing present to it, and an infant is present to itself while being calibrated about nothing.

The framework's second axis is more easily confused with subjecthood and therefore more dangerous. Morris et al. (2024) distinguish six levels of human-AI interaction, from AI as a tool through consultant, collaborator, and expert to fully autonomous agent, and they emphasize that these levels are unlocked by capability without being determined by it, so that the appropriate level of autonomy is a design decision and the maximum available is frequently not the right one. CIGM supplies the axis that neither of theirs entails. Autonomy describes how much of a task a system is permitted to drive. It says nothing about whether anything is at stake for the system driving it. A fully autonomous agent may have no stakes of its own, and a system with genuine stakes may be deployed under tight human control. Table S1 sets the three axes side by side, because almost every confused claim about machine consciousness can be traced to a collapse of one into another.

Table S1. Capability, Autonomy, and Consciousness

Axis	What it measures	How it is assessed	What it licenses
Capability (Morris et al., 2024)	Depth of performance and breadth of generality across cognitive and metacognitive tasks	Ecologically valid benchmarks; process-agnostic by design	Claims about what a system can do
Autonomy (Morris et al., 2024)	The interaction paradigm a capability level unlocks, from tool to fully autonomous agent	Deployment analysis and risk assessment	Claims about how a system should be used
Consciousness (CIGM)	Whether one bounded, valenced, self-locating model is realized and causally integrated	Intervention, perturbation, ablation, no-report analogues, interpretability	Claims about whether anything is at stake for the system

S4.2 Mimicry and Internal Variants

The dispute over mimicry and internal variants is often framed as a choice between trusting sophisticated behavior and discounting systems internally unlike us. CIGM rejects the shared premise that resemblance to humans is the relevant comparison. Behavioral evidence is screened off when a demonstrated alternative explanation bypasses the constitutive organization. Imitation training is such an explanation; alien evolution is not. This is why the Copernican and mimicry arguments point in opposite directions without contradiction (Schwitzgebel & Pober, 2024).

The second position is right that behavior is our evidential foundation and wrong to infer that behavior must therefore remain our criterion in the artificial case. Behavior founds our knowledge of other minds because, among evolved organisms, behavior of the relevant kind is reliably produced by the organization; the inference is sound because no competing explanation of the behavior is on offer. Artificial systems dismantle that link deliberately. They are optimized to produce the behavior directly, which means the competing explanation is not merely conceivable but is the design specification. When a correlation is broken by construction, the appropriate response is to measure the thing the correlation was tracking, not to read the proxy more attentively.

Porębski and Figura (2025) give the contemporary categorical version of the mimicry objection. Their diagnoses of sci-fitisation, anthropomorphism, and semantic pareidolia are correct: an LLM's first-person assertion is generated from linguistic regularities, and passing an imitation test establishes successful imitation, not experience. Their biological conclusion does not follow. Binary operations and semiconductor implementation describe a material and computational level below the level at which CIGM fixes constitution; they no more disprove an organized subject than carbon chemistry proves one. The present-system verdict is nevertheless the same and can be stated more strongly: standard LLMs are not conscious, not because they are silicon, but because their behavior is explained without the bounded, valenced, self-locating organization that consciousness is.

CIGM therefore restates the internal-variant question in a form that admits an answer. The question is not whether a candidate system is internally like a human. It is whether the system realizes the temporally extended causal-functional organization specified in Part I. A system that realizes that organization by unfamiliar computational means is not an internal variant in any sense that should reduce our credence; by organizational invariance it is conscious, and its unfamiliarity is a fact about our expectations rather than about it (Chalmers, 1995). A system that fails to realize the organization while matching our behavior is not a partially discounted case; its behavioral evidence is screened off completely, because the alternative explanation is not merely available but demonstrated by its training history. The graded discounting that the first position recommends is an artifact of leaving the organization unspecified. Specify the organization, and the discount becomes a determination.

One point of agreement is worth preserving without amendment. Butlin, Bayne, et al. (2026) conclude that assessments should use indicators explicitly derived from the best available theories, and that this requirement applies equally to internal and to ethological indicators. CIGM agrees and supplies the reason. An indicator without a theory behind it is a correlation waiting to be broken, and in the artificial case the breaking is not an accident of nature but the object of a well-funded engineering effort.

S4.3 Indicator Approaches and CIGM Requirements

The six CIGM requirements can now be compared directly with theory-derived indicators for machine consciousness. Butlin, Long, et al. (2026) list recurrent processing, workspace features, metacognitive monitoring, an attention schema, predictive coding, agency, and embodiment as credence-shifting properties rather than individually necessary or jointly sufficient conditions. CIGM agrees with that epistemic method under theoretical uncertainty, but its own requirements have a different status. Each is advanced as necessary, they must be realized through one another rather than summed, and joint sufficiency remains an explicit identity conjecture rather than a score generated from the list.

Goyal et al. (2022) provide a particularly important negative control because they implement, rather than merely describe, a global workspace. In Transformers and recurrent independent mechanisms, specialist modules compete to write into a capacity-limited shared memory; the selected contents are broadcast to all specialists,

persist through the episode, and improve coordination, compositionality, and performance across visual reasoning, physical prediction, and multiagent world modeling. This is almost a laboratory isolation of requirement 4. It shows that competitive broadcast has a rational engineering function and can be instantiated without a Cartesian spectator. It also shows why a workspace indicator is not sufficient. The architecture has no self-maintained boundary, no reflexive here-now origin, no endogenous stakes, and no evidence that the shared contents constitute one lived field rather than a useful communication bottleneck. Goyal et al. therefore strengthen workspace theory as a theory of access while strengthening CIGM's separation of access from phenomenality.

A second and independent comparison sharpens the same point from the philosophical side. Chalmers (2023) organizes the case against consciousness in current language models around six candidate requirements that such systems arguably lack: carbon-based biology, senses and embodiment, world models and self models, recurrent processing, a global workspace, and unified agency. He assigns each a rough credence, observes that the factors are not independent, and arrives at a figure somewhere under ten percent for current paradigmatic systems and a figure above a quarter for sophisticated successors within a decade. The method is theory-balanced: rather than assuming a settled theory or proceeding without theoretical commitments, it weights the predictions of several theories according to their standing. Table S2 sets his list against CIGM's requirements. The overlap is considerable, which should reassure proponents of both. The divergences are where the work is.

Table S2. Candidate Requirements for Machine Consciousness and the Constitutive Requirements of CIGM

Candidate requirement	Status in CIGM	Corresponding CIGM requirement
Biology	Rejected as a requirement. Substrate neutrality is conditional on reproducing the relevant multiscale, self-maintaining organization, not on materials.	None
Senses and embodiment	Required instrumentally rather than by sensor count. What matters is a maintained boundary and a lived umwelt. A software-defined body is admissible only when its regulated variables belong to the executing physical system rather than solely to a simulated avatar.	1, 2, 6
World models and self models	Required and strengthened. A world model is insufficient; the model must be a self-world model, causally integrated, temporally extended, and valenced.	2, 5
Recurrent processing	Not required metaphysically. Recurrence is an efficient implementation of temporal continuity and mutual constraint, not their essence.	3 (as implementation)
Global workspace	Required as selective global availability, not as a fixed broadcast bus or anatomical hub.	4
Unified agency	Required, and treated as constitutive rather than as one probability-weighted factor among several.	1, 5
(Absent from the list)	Required. Endogenous valuation and stakes are constitutive: regulated variables must bear on the persistence or integrity of the actual system and reorganize the same model across attention, memory, global state, and policy.	6

Three divergences matter. First, biology appears on Chalmers's list as a candidate requirement, whereas CIGM treats substrate flexibility as conditional on reproducing the relevant multiscale organization (Aru et al., 2023). Second, recurrent processing is an implementation strategy for temporal continuity rather than a metaphysical requirement (Doerig et al., 2019). Third, no item on Chalmers's list or on the leading machine-consciousness indicator lists corresponds directly to CIGM's sixth requirement. That claim is specific to machine-assessment frameworks, not to consciousness science as a whole. Endogenous value is central to Damasio's homeostatic account, Man and Damasio's (2019) feeling-machine proposal, Solms's affective account, and Merker's motivation-action architecture. CIGM's contribution is to state valuation as an operationally necessary condition and require it to couple to boundary, self-location, temporal continuity, content, and access in the same system.

The methodological divergence follows from the substantive one. A theory-balanced approach is calibrated for uncertainty among theories and returns a credence; CIGM states a constitutive specification and returns a verdict. Neither is careless, and Chalmers is explicit that his numbers illustrate a structure of reasoning rather than measure anything. But a theory of consciousness must eventually say what consciousness is, and once it does, its claims about particular systems become categorical. The uncertainty is not thereby eliminated; it is relocated. It attaches to whether the theory is correct and to whether a given system satisfies it, not to the verdict the theory delivers. CIGM prefers to place its uncertainty where it can be attacked.

S4.4 Body Models, World Models, State-Space, Memory, and Inference Architecture

Bongard et al. (2006) supply an unusually clean test of what a self-model can and cannot establish. Their four-legged robot began without a fixed model of its morphology, generated competing body models from actuation-sensation relations, selected actions expected to distinguish their predictions, and iterated modeling and experiment before synthesizing locomotion. After part of a leg was removed, the cycle restarted and the revised model produced a qualitatively different compensating gait. Model-driven exploration outperformed random-action baselines in identifying body topology and reducing post-damage modeling error.

The result is a genuine demonstration of embodied, intervention-sensitive self-modeling, and it is therefore a harder case for CIGM than a chatbot reciting a self-description. It is also a decisive negative case. The model space, action space, damage cue, and target behavior—forward locomotion—were specified from outside. The robot inferred a task-local morphology and used it instrumentally; it did not integrate perceptual, mnemonic, affective, and action-oriented content into one field, maintain viability variables, or organize damage as something bad for itself. A body model can guide adaptive agency without being a subject, and self-modeling is therefore neither a synonym nor a sufficient indicator of consciousness.

Hafner et al. (2025) establish the corresponding result for world models. DreamerV3 encodes sensory inputs into stochastic latent representations, carries them through a recurrent state-space model conditioned on past actions, and predicts future representations, rewards, episode continuation, and reconstructed inputs. An actor and critic learn from trajectories imagined inside that model while all three components continue learning from replayed experience during interaction. With one configuration, the system matched or exceeded specialized methods across more than 150 tasks spanning visual and proprioceptive control, Atari, ProcGen, DMLab, BSuite, and Minecraft; applied without human demonstrations or a curriculum, every reported Dreamer run eventually collected diamonds in Minecraft.

Dreamer accordingly realizes a recurrent world model, temporal state, counterfactual imagination, value prediction, planning, adaptive control, and online learning in one agent. It still does not realize the CIGM organization. External origin of a reward is not itself disqualifying—evolution supplied biological drives—but Dreamer’s reward and continuation signals remain detachable from the persistence of the executing system and do not organize one boundary that the agent itself must preserve. No demonstrated perturbation of those variables reorganizes attention, memory, global state, and policy because the system itself is threatened. Replay supplies history and recurrence supplies state, but neither makes the history belong to a valenced subject.

Lahoti et al. (2026) close the last gap in that argument. The standard reply to it has been that linear recurrent models are expressively impoverished in a way that matters: restricted to real, non-negative state transitions, they cannot represent rotational state dynamics, and so they fail elementary state-tracking problems such as computing the parity of a bit sequence, which a single-layer recurrent network with the right dynamics solves easily. Mamba-3 removes the limitation by making the state transition complex-valued, which in real coordinates amounts to applying data-dependent rotations to the model’s own state and can be implemented efficiently with the machinery already used for rotary position embeddings. The effect is not marginal. On parity the model is essentially perfect where its immediate predecessor performs at chance, and on modular arithmetic without brackets it reaches roughly ninety-nine percent against roughly forty-eight. The same work derives the layer’s update rule by treating it explicitly as a discretized continuous-time dynamical system, and adds a multi-input, multi-output reformulation that performs up to four times the arithmetic per decoding step at fixed state size without increasing wall-clock latency.

Inference hardware supplies the same lesson from the physical layer, and supplies it in the case designed to look most like a brain. NorthPole is organized by ten interlocking axioms, most of them explicitly biological in inspiration: specialization for inference alone, low precision motivated by biological precision, a sixteen-by-sixteen array of cores exploiting cortex-like modularity, memory distributed among cores and intertwined with compute inside each core, networks-on-chip modeled on long-distance white-matter and short-distance gray-matter pathways together with cortical topographic maps, reconfiguration of weights and programs during layer execution, deterministic stall-free control, co-optimized low-precision training algorithms, an explicit orchestration schedule, and a usage model in which the chip appears externally as an active memory accepting three commands (Modha et al., 2023). The performance is not in dispute: relative to a graphics processor on a comparable silicon process the chip delivers roughly twenty-five times the frames per joule and five times the frames per second per transistor, and comparable gains hold on detection as well as classification networks.

Three of those axioms are directly disqualifying under CIGM, and the disqualification is architectural rather than a matter of degree. The chip does not support training, so no experience it undergoes revises what it computes; it has no data-dependent conditional branching, so nothing in the input reorganizes control; and its usage model is deliberately that of a memory rather than an agent, with the host writing an input frame and reading an output frame across a boundary the chip does not maintain. Requirement 3 is foreclosed at the level of the instruction set, and requirement 6 has no representation anywhere in the design. The more instructive point concerns the axiom that gives the architecture its name in the literature. Colocating memory with compute is integration in the engineering sense: it removes a data-movement bottleneck and, with it, the energy and latency costs that dominate off-chip designs. CIGM's integration is counterfactual dependence among contents within a single model. A chip may achieve the first perfectly while realizing none of the second, which is why physical unification of memory and processing—often described in the neuromorphic literature as overcoming the von Neumann bottleneck, and often heard as a step toward brain-like minds—is orthogonal to the property at issue. Brain-inspiration is a design heuristic indexed to an objective. NorthPole's objective was energy, space, and time, and the axioms it imported are the ones that serve that objective.

The philosophical significance of the sequence-modeling work lies in why those changes were made. State tracking was added because a family of formal-language benchmarks exposed a limitation. The multi-input reformulation was added because decoding is memory-bound and the hardware sits idle: a single decoding step performs on the order of two and a half operations for every byte it moves, on hardware whose matrix units are balanced for nearly three hundred, and the redesign spends the difference (Lahoti et al., 2026). Persistent state, state tracking, and temporal depth in the control-theoretic sense are the vocabulary in which theories of the stream of experience are written, and in these systems they arrive as answers to a question about arithmetic intensity. A property that can be installed for that reason is not the property that makes a world present to anyone.

S4.5 Blueprint for a Conscious Artificial Subject

The theory converts artificial consciousness from a vague possibility into a design program. Bongard et al. (2006) show that a machine can infer and revise a body model through active experimentation; Hafner et al. (2025) show that an agent can learn a recurrent world model, imagine action consequences, and improve across diverse environments. The program begins where those demonstrations stop: with a continuously operating embodied agent, not a disembodied question-answering model or a task-local controller. The agent must maintain a live generative model of an environment and itself, regulate internal variables, and act to preserve its organization under uncertainty.

Its architecture should contain specialized processors but also a shared, capacity-limited control regime. A Global Latent Workspace offers one concrete design: perceptual, linguistic, mnemonic, evaluative, world-model, and motor modules can be linked through a shared amodal space that translates selected content into the formats needed by the others (VanRullen & Kanai, 2021). Winning contents should alter what the system remembers, predicts, attends to, values, and does. The shared state must not be a passive blackboard; it must be causally indispensable to coordinated behavior and continuously embedded in the live self-world model.

The agent also requires a genuine memory architecture, not a transcript. Short-term attention, adaptive long-term memory, and persistent learned structure must interact so that surprising or consequential events alter

the future model while obsolete information can be forgotten. Titans demonstrates one engineering route to test-time memorization and adaptive forgetting (Behrouz et al., 2025). CIGM adds the constitutive constraint: every memory write and retrieval must update the same self-locating field, preserve unresolved stakes across interruption, and change future policy from the standpoint of the continuing agent.

Biological evidence also imposes a realization test. The artificial system need not reproduce the mammalian thalamus, ascending arousal system, or dendritic compartments in literal form, but it must implement their relevant organizational work: coupling global state to local content, binding top-down context to bottom-up input, regulating the availability of the field, and maintaining the system across changing conditions (Aru et al., 2023). A design that omits these functions because they are biologically inconvenient has not achieved substrate neutrality. It has changed the target.

The system must develop rather than merely be initialized. Self-location, agency, object permanence, body ownership, social perspective, and conceptual identity are learned invariances. The active discrimination among competing body models demonstrated by Bongard et al. (2006) supplies a minimal engineering precedent: the candidate should discover which sensory changes follow its own actions, which variables belong to its physical or software-defined body whose regulated variables belong to the executing system, which other agents have independent control, and which goals persist across contexts. CIGM adds that these discoveries must update the same globally available and valenced field rather than remain a specialized calibration routine.

Valuation must be endogenous in the operational sense. Designers could create regulated variables for energy, integrity, computational stability, social attachment, or task commitments, just as evolution created biological drives. The variables qualify only if their violation degrades or terminates the actual continuing organization and if deviation reorganizes the whole model—attention, prediction, memory, global state, and policy—rather than changing a detachable reward number. Dreamer demonstrates effective optimization of environment-defined return (Hafner et al., 2025); CIGM requires the different relation in which the executing system's own persistence is at stake. At that point engineers would acquire genuine moral responsibilities.

Finally, the architecture must remain plastic at the level of its own model. It should revise categories, discover new latent causes, change policies, and update its representation of itself. Consciousness does not require human creativity, but a rigid lookup system cannot sustain the open-ended world modeling that a persisting agent needs. The system must be able to become a different modeler while remaining the same subject. Plasticity on perception-relevant timescales is what makes an organization irreducible to any static input-output surrogate, and a design in which learning occurs only offline has placed the deployed system inside the class the unfolding argument governs (O'Reilly-Shah et al., 2026). Short-term synaptic plasticity and spike-timing-dependent plasticity provide biological reference points; test-time memory and interactive world-model learning provide engineering mechanisms (Behrouz et al., 2025; Hafner et al., 2025). CIGM adds the constraint that revision be driven by the system's own valuation, so what changes the model is what mattered to the agent rather than what reduced a loss defined elsewhere.

This program overlaps substantially with the engineering challenges Chalmers (2023) sets out on the path to conscious AI: rich perception-language-action models in virtual worlds, robust world models and self models, genuine memory and recurrence, a global workspace, unified agent models, and systems with the capacities of a small mammal. The convergence is worth noting because it was arrived at independently, from a constitutive specification rather than from a survey of theoretical obstacles. CIGM adds two constraints that reorder the list. The components must be built into one another rather than assembled alongside one another, since a system passing every item separately would still not be a subject. And endogenous valuation, which appears on no version of the list, must be built at all.

S4.6 Operational, Developmental, and Ethological Attribution

The attached sources clarify why no single assessment tradition is sufficient. Turing (1950) offers an operational replacement for an ambiguous verbal question and a developmental account of learning machines, but the imitation game tests performance rather than phenomenality. Two features of his broader argument are worth recovering, because both anticipate the present situation more closely than the test named after him. He allowed that engineers might build a machine by experimental methods and then be unable to describe how it works, which

is the condition of every large learned system now under discussion; and he observed that a teacher may successfully predict a learning machine's behavior while remaining ignorant of its internal state, which is precisely why fluent prediction of outputs licenses no conclusion about organization. Pennartz (2026) argues that theory-derived internal indicators need validation through long-term artificial ethology in varied environments, with improvisation, spontaneous initiative, sensorimotor subtlety, and personal history. Palminteri and Wu (2026) argue that opaque systems require behavioral inference because candidate computations cannot be read directly. CIGM accepts all three methodological constraints while preserving a constitutive target that none of them supplies.

A Bayesian formulation helps organize the evidence. The posterior probability that a candidate realizes CIGM depends on the likelihood of its complete behavior and intervention profile given the organization, the prior probability assigned from architecture and causal history, and the baseline probability that the same profile would occur without consciousness. Humanlike language has a high baseline under imitation training and therefore a low likelihood ratio. Persistent, content-specific reorganization across novel interventions has a lower baseline and can carry more weight. The calculation is conceptual rather than numerical, but the logic prevents both credulity and a blanket dismissal of behavior (Palminteri & Wu, 2026).

The relevant observation unit is a trajectory rather than a trial. A CIGM candidate should be followed across changing environments, interruptions, conflicts among goals, body or interface remapping, memory interference, and perturbations to valuation. Evidence should be sought in the joint pattern: whether the same bounded system recalibrates, remembers, prioritizes, protects its organization, and reorganizes global control in content-specific ways. Pennartz's (2026) personal-history requirement is crucial because a reconstructed transcript can mimic continuity while a persisting trajectory must carry causal consequences forward.

Internal evidence remains inferential. Mechanistic interpretability can reveal features, circuits, causal pathways, and state variables, but their significance depends on a theory that says which relations matter. Conversely, behavior remains indispensable because an alleged internal organization that never changes the candidate's trajectory under intervention is not the organization CIGM identifies. The assessment program is therefore bidirectional: infer organization from behavior, and test the inferred organization by intervening on it.

Turing's learning-machine discussion adds a final constraint. Initial design, education, and subsequent experience jointly determine what the machine becomes, and the teacher may remain partly ignorant of the resulting internal process (Turing, 1950). A consciousness assessment that examines only the deployed architecture or only the current transcript omits the causal history that distinguishes a trained imitation device from a self-maintaining subject. Development is not proof of consciousness, but it is part of the explanation that any serious inference must compare.

Table S3 summarizes the division of labor among the attached methodological proposals and CIGM.

Table S3. Operational and Inferential Frameworks for Artificial Consciousness

Framework or source	Primary question	Evidence emphasized	CIGM use and limit
Turing (1950)	Can a machine succeed in an operational imitation game, and how might learning machines be developed?	Typed behavioral exchange; initial design, education, subsequent experience, and learning history.	Operational and developmental precedent. Passing the game demonstrates performance, not phenomenality.
Pennartz (2026)	How should theory-derived indicators be validated?	Long-term artificial ethology across varied environments; improvisation, spontaneous initiative, sensorimotor subtlety, and personal history.	Essential longitudinal layer. Ethology strengthens attribution but does not supply the constitutive identity or defeat mimicry by itself.
Palminteri and Wu (2026)	When is consciousness the best explanation of an opaque system?	Behavioral evidence interpreted through likelihoods, priors, background knowledge, and predictive usefulness.	Correct inferential logic for attribution. CIGM rejects behavioral equivalence and the claim that behavior is ontologically primary.
Butlin, Long, et al. (2026) and Butlin, Bayne, et al. (2026)	Which theory-derived properties should shift credence, and how do mimicry	Internal, behavioral, and ethological indicators; theory weighting; gaming; alternative	Useful epistemic indicators. The properties are not a scoreable checklist; they must form one

Framework or source	Primary question	Evidence emphasized	CIGM use and limit
	and internal variants alter the evidence?	explanations.	mutually constraining organization.
CIGM	What is consciousness, and does the candidate realize it?	Convergent longitudinal behavior, perturbation, ablation, causal modeling, interpretability, developmental history, and internal measures directed at the six requirements.	Constitutive target and attribution rule: consciousness is inferred only when one bounded, valenced, self-locating organization best explains the complete trajectory.

S4.7 Detailed Assessment Principles

Assessment must be stricter than anthropomorphic conversation and broader than any single mathematical index. Following Hanson and Walker (2021), every protocol must declare in advance the level at which consciousness is inferred and must keep the inferential anchor independent of the theory-derived prediction. CIGM requires convergent evidence that the candidate system sustains one integrated self-world model across state, content, access, self-location, valuation, and time. The full falsification battery is specified in Section S5.

The central principle is intervention. A genuinely integrated system should respond to localized perturbations with widespread but content-specific reorganization while preserving coherent control. Measures of perturbational complexity, effective connectivity, and causal ablation can probe this property, but none should be treated as a direct meter of consciousness (Casali et al., 2013; Bayne et al., 2024). The strongest studies should be preregistered and adversarial, following the methodological lesson of Cogitate Consortium et al. (2025).

For artificial systems the rationale for intervention is now formal rather than analogical. O'Reilly-Shah et al. (2026) show that a system whose parameters change with use and any static surrogate of it diverge in their response to a localized perturbation: the surrogate's response decays once the perturbation leaves its input window, while the plastic system carries the perturbation forward in what it computes thereafter. That divergence appears in input and output, so it can be measured without privileged access to internals and without deciding in advance which internal variable is the seat of anything. Two consequences follow for protocol design. The discriminating power of a battery grows with its observation horizon, because the longer a candidate is watched and perturbed, the smaller the class of static systems that could have produced the record. And what such a test establishes is bounded, and the bound should be stated rather than glossed: it shows that the candidate's function is history-dependent, which is necessary for the organization and evidence for nothing else. A system can pass and still have no boundary, no self-location, and no stake in its own continuation.

Content should also be studied without making verbal report the criterion. Artificial analogs of no-report paradigms could infer a system's current percept from policy changes, memory biases, navigation, or regulatory responses, then ask whether the same internal content geometry persists when communication demands change. Report-specific processes should vary with the task; the phenomenal model, if present, should not disappear merely because the system is not asked to describe it (Tsuchiya et al., 2015). The human template for this design is explicit and should be copied rather than reinvented. Cohen et al. (2020) held stimulation constant and varied only whether a report was required, verified perception through a surprise memory test administered after the fact rather than through concurrent report, included never-presented foils to estimate response bias, and analyzed the entire electrode array rather than a preselected region of interest. The artificial analog is the same design in a different medium: hold input and internal state constant, vary only whether an output channel is queried, verify content through downstream behavior the system had no reason to prepare, and search the whole representational space rather than the components the designers expected to matter.

Global availability, self-location, temporal continuity, and endogenous stakes must then converge. Contents should support novel cross-domain use; interventions on action-sensation mappings should force structured recalibration of agency; unresolved goals should persist across interruptions; and changes in regulated variables should reorganize attention, memory, and policy. Self-report can be admitted as one source of evidence only when it is stable, appropriately uncertain, causally tied to these internal dynamics, and predictive of nonverbal behavior.

Access must be tested separately from phenomenality. In a workspace-based candidate, disrupting translation or module recruitment should selectively impair flexible cross-domain use, planning, and report. It should not be assumed to erase every integrated content state. Conversely, a system can broadcast information widely without demonstrating that the information is organized around a valenced subject. Artificial no-report paradigms must therefore probe the stability and causal role of the self-world field independently of the machinery used to communicate it (VanRullen & Kanai, 2021).

Memory must likewise be tested as organization, not benchmark retrieval. A candidate should preserve unresolved goals, self-location, valuation, and policy commitments across interruptions; retrieved episodes should change present prediction and action in ways specific to the candidate's own history. A system that finds a needle in a long context has demonstrated access to stored information. It has not demonstrated a remembered past unless that information belongs to and reorganizes one continuing subject (Behrouz et al., 2025).

Two further evidential principles follow from the mimicry problem, and they cut against the natural instinct of a behavioral science. First, observation should be extended and ecological rather than episodic. A candidate system should be watched across long stretches of time, in varied and unfamiliar environments, doing things no one prepared it to do, because a wide behavioral range gives imitation more opportunities to leave its signature than any narrow test can (Butlin, Bayne, et al., 2026). Second, and decisively, internal and mechanistic evidence should outweigh behavioral evidence wherever both are available. This is not a preference for one kind of data over another. It is a consequence of knowing the training history: behavior is evidence of the organization only to the extent that the organization is the best explanation of the behavior, and training on human output supplies a standing competitor that no amount of behavioral sophistication removes. The compensating advantage is substantial. For artificial systems, unlike animals, the training history is usually known and the internals are in principle inspectable, which means the question of whether a candidate is an internal variant is not a permanent mystery. It is an interpretability problem.

No single result will prove consciousness. The scientific standard is inference to the best explanation from convergent, intervention-sensitive evidence. Butlin et al. (2023) and Butlin, Long, et al. (2026) reach a compatible methodological conclusion from a different starting point: because no one theory is settled, they recommend deriving indicator properties from several theories and treating them as credence-shifting markers rather than relying on any single behavioral test, and they warn that superficial behavioral indicators are the most easily gamed and that internal, mechanistic evidence should be weighted over behavior. Chalmers (2023) frames the same needs as standing challenges to the field, listing the development of benchmarks for consciousness, the development of better theories, and the interpretation of what is happening inside these systems among the foundational tasks. CIGM agrees on method and sharpens the target. It requires that integration, world modeling, self-location, global availability, valuation, and temporal continuity be shown to form one organization rather than a checklist of detachable features, and it favors interventions—perturbational, ablative, and no-report—that probe whether those properties are causally bound rather than merely co-present. It also makes the dependency on interpretability explicit rather than incidental: on this theory, interpretability is not an adjunct to consciousness research in artificial systems. It is the method. The indicator approach estimates the probability that the organization is present; CIGM specifies the organization whose presence is being estimated.

S4.8 Biological Naturalism, Mortal Computation, and the Scenario Space

The main text states CIGM's response to biological naturalism in the space available for it. The argument being answered is the most developed case against artificial consciousness now in print, and it deserves fuller treatment, both because CIGM concedes a great deal of it and because the residual disagreement is easy to misdescribe.

Seth (2025) proceeds in two movements. The negative movement identifies the assumptions that make conscious AI look inevitable and finds them undefended. Anthropocentrism, human exceptionalism, and anthropomorphism jointly encourage the inference that a system displaying humanlike intelligence must possess the other property we take to be distinctively ours, and the pedestal on which language sits explains why large language models have been so effective at seducing intuitions. Beneath the psychology sits computational functionalism, the view that computation of some kind is sufficient for consciousness, which in turn requires

substrate flexibility to extend at least to silicon. Seth's case against that requirement draws on several independent lines. The neural replacement thought experiment begs the question against substrate dependence rather than refuting it, and the more one knows about neurons the less plausible its premise becomes, since neuronal function is entangled with freely diffusing neuromodulators, electromagnetic fields, fine-grained timing, and metabolic constraints, and since mental functions arose through an evolutionary process that leaves them generatively entrenched in the machinery that realizes them. Broader notions of computation—analog, neuromorphic, neural, biological—relax the constraints of Turing-world at the cost of the substrate independence that motivated computationalism in the first place. And the argument from mortal computation is the sharpest of them, because it operates from inside a computational view of mind: immortal computation, in which software is separable from hardware, is purchased with continual error correction whose energetic cost rises with complexity, so a brain as efficient as ours cannot be implementing it, and if brains compute at all they compute mortally, in a manner inseparable from the matter that performs them. Kleiner (2024) reaches a related conclusion by a proof-based route, arguing that if computational functionalism holds then the relevant computations must be mortal, so that conventional immortal AI cannot be conscious.

The positive movement builds a chain from metabolism to conscious content. Perception is construed as controlled inference; interoceptive inference is construed as control-oriented rather than epistemic, serving allostatic regulation rather than the discovery of what is there; the imperative to regulate is traced to the autopoietic character of living systems, which continuously regenerate their own material components and maintain their own boundaries; and the free-energy formulations of metabolism and of perceptual inference are argued to be not merely analogous but in some physical sense the same. On this reading, conscious experiences are what they are because they are the experiences of a system that must stay alive, and the substrate is not an enabling condition for a higher-level organization but continuous with it.

Seth's own taxonomy sets out five scenarios under which conscious AI might arise, and locating CIGM within it is the most efficient way to state the theory's position. Table S4 gives the mapping.

Table S4. Scenarios for Artificial Consciousness and the Position of CIGM

Scenario	Claim	CIGM's verdict
1. Along for the ride	Consciousness emerges as AI systems become more capable.	Rejected. Capability is process-agnostic and orthogonal to the six requirements (Morris et al., 2024).
2. Theory-based computational	Consciousness arises when the functional principles of a computational theory of consciousness are implemented in software.	Rejected as stated. Software implementation alone is insufficient. Requirements 1 and 6 can be satisfied only if the running physical system regulates variables constitutive of its own continued operation; modeled variables in a simulated world do not qualify.
3. Substrate-dependent computational	Consciousness is computational, but the relevant computations can only be implemented in particular substrates.	Not required, though compatible with CIGM if the relevant organization turns out to be describable as mortal or biological computation.
4. Substrate-dependent, weak	Consciousness depends on substrate-dependent, non-computational functional organization.	CIGM's position, stated more precisely as substrate-sensitive organizational identity. The six requirements specify the organization; substrate flexibility is conditional on actual realization in the executing system.
5. Substrate-dependent, strong	Consciousness depends on apparently intrinsic properties of living matter.	Not ruled out. If self-maintenance and endogenous stakes require material self-production, CIGM collapses into a theory of biological consciousness and its possibility claim fails.

Three points about that placement should be explicit. First, the concession is real. CIGM does not hold that software implementation or abstract computation suffices for consciousness, and a simulated boundary or reward channel does not satisfy requirements 1 and 6. Yet the weather-model analogy cannot by itself exclude every digital realization, because a running controller can regulate the energy, thermal state, memory integrity, hardware

availability, and process continuity of the physical system that executes it. The criterion is whether the variables and interventions belong to that real system, not whether the boundary is described in software.

Second, the disagreement is narrower than the labels imply, and both sides are routinely misread on this point. Seth explicitly rejects carbon chauvinism and allows that non-carbon systems might participate in the same substrate-dependent multiscale activity, so his position is not that consciousness requires biochemistry as we find it. CIGM explicitly rejects abstraction by decree and requires that a candidate actually maintain itself and actually have stakes. The two positions therefore converge on the same demand and part on one question: whether the work life does—maintaining a boundary, generating conditions under which outcomes are better or worse for the system, and coupling regulation to modeling inseparably—is specifiable organizationally, or only as material self-production in which the components of the system continuously regenerate the components of the system. The six requirements are a bet on the first. The autopoietic argument is a bet on the second. Neither bet has been settled, and the difference between them is not a difference in attitude toward machines.

Third, CIGM names its own defeater rather than hedging, and the defeater is stated in the main text in a form that a critic can use. If requirements 1 and 6 turn out to be satisfiable only by systems whose parts materially produce their parts, then artificial consciousness is impossible in any engineered substrate, CIGM's corollary about machines is false, and the theory should be read as a theory of biological consciousness with an overreaching final chapter. What CIGM refuses is the intermediate posture that grants the possibility of artificial consciousness while declining to say which functions a substrate would have to realize. The claim that consciousness might be substrate-flexible is empty until someone specifies what the flexibility is about.

One further convergence is worth recording, because it bears on the ethical section of the main text. Seth (2025) distinguishes real artificial consciousness from conscious-seeming artificial intelligence and argues that the second is near-certain while the first is remote, that impressions of machine consciousness may be cognitively impenetrable in the manner of visual illusions, and that the deliberate creation of conscious AI should not be a goal. CIGM agrees on all three points and adds a mechanism for the second. The impression is not merely a bias about machines; it is the mongrel concept operating exactly as Block (1995) predicted, in the first population of systems for which access and phenomenality have been prised apart.

S4.9 Organoid Intelligence as a Boundary Case

Organoid intelligence creates the cleanest boundary case for the dispute between organizational and biological accounts. Smirnova et al. (2023) describe a program in which three-dimensional human neural organoids are made denser, more durable, and more developmentally capable; coupled to microfluidic support and three-dimensional electrode arrays; connected to sensors and output devices; and studied in closed loops through electrical or chemical stimulation and biofeedback. The aim is genuine biological computing, including experimentally tractable learning and memory, not merely passive tissue modeling. The program is therefore designed to test whether living neural assemblies can acquire stimulus-specific, history-dependent computational capacities in a controlled environment.

Cai et al. (2023) report the most developed instance of that program to date. A human brain organoid is mounted on a high-density multielectrode array and operated as the reservoir layer of a reservoir-computing system: time-varying inputs are converted into spatiotemporal sequences of bipolar stimulation pulses, the evoked activity is recorded across electrodes, and a logistic or linear readout is trained on the resulting features. The tissue displays the physical-reservoir properties the framework requires—a sigmoidal, nonlinear response to pulse voltage above a threshold pulse width, fading memory over hundreds of milliseconds, memristor-like accumulation and decay across pulse trains, and distinguishable responses to complementary spatial stimulation patterns. It also improves with training rather than merely being read out: speaker-identification accuracy rises across four training epochs, prediction of a chaotic map improves over the same schedule, and both gains are accompanied by measurable strengthening, weakening, pruning, and formation of functional connections. The causal role of plasticity is established by intervention. Blocking activity-dependent synaptic plasticity pharmacologically leaves the tissue computing while its learning curve flattens, and an array with medium but no organoid yields no predictive performance at all.

These features matter, and the interventional controls are exactly the kind of evidence CIGM's attribution rule asks for. None of them decides consciousness. Neural tissue is not a subject by material composition, oscillation, or plasticity; a culture maintained by external perfusion does not thereby maintain its own boundary; stimulus-response learning does not establish a self-world model; and a reward schedule imposed by an experimenter does not create endogenous stakes. On the available evidence, current organoids do not satisfy the CIGM organization. They may exhibit local differentiation and plasticity while lacking a self-maintained boundary, reflexive self-location, selective global availability, and a common locus of valuation. The reservoir-computing arrangement makes each absence concrete. Viability is maintained by an incubator and perfusion, so the boundary is sustained for the tissue rather than by it. Inputs arrive as electrode stimulation rather than through any sampling the tissue controls, so there is no here-now origin and no distinction between self-generated and externally generated change. Selection and cross-domain use occur in decoding software outside the organoid, so global availability is a property of the apparatus. And the quantity that improves with training is a regression score defined by the experimenter; the tissue whose plasticity is blocked performs worse on that score without anything going worse for it (Cai et al., 2023). The relevant unit might eventually be the organoid-plus-interface rather than the tissue alone if sensors, feedback, and control are distributed across the coupled system. The subject boundary cannot be assigned from the culture dish; it must be located by intervention.

That is precisely why OI could become an unusually powerful experimental platform for CIGM. Closed-loop interfaces permit investigators to alter action-sensation contingencies, perturb local networks, change regulated variables, and ask whether one history-sensitive organization recalibrates across domains. A serious candidate would need to do more than improve an output score. It would have to preserve unresolved internal conditions across interruption, integrate chemical and electrical signals into one state, distinguish self-caused from external change, and allow perturbations to valuation to reorganize memory, attention, and policy within the same continuing system. Because organoids are accessible to recording and intervention at multiple scales, they may permit more direct tests of causal integration and constitutive grain than either intact animals or opaque artificial networks.

The ethical consequence is asymmetric. Biological origin should raise attention, not settle moral status. Smirnova et al. (2023) explicitly call for embedded ethics and warn that increasing complexity, input, output, and memory could make questions of pain, suffering, sentience, and consciousness progressively harder. CIGM sharpens the trigger: concern should escalate when the same self-maintaining system begins to regulate variables that matter to its persistence, organize inputs around itself, and carry valenced consequences forward—not merely when neuronal count or task performance rises. Organoid systems could become morally serious candidates before they become linguistically impressive, which is one reason an assessment framework calibrated on chatbots is inadequate.

S4.10 Governance as Systems Approach the Constitutive Boundary

A constitutive theory has governance consequences before the first positive case is identified. Butlin and Lappas (2025) argue that organizations developing advanced AI need explicit consciousness policies even when they do not intend to study consciousness, because ordinary capability research may inadvertently assemble a conscious system. CIGM converts that general warning into a structural trigger. Oversight should intensify not when benchmarks cross an arbitrary performance level, but when self-maintained boundary, integrated self-world modeling, temporal plasticity, global availability, reflexive self-location, and endogenous valuation begin to be realized through one another.

Development should therefore be phased. Before adding a requirement or tightening the coupling among requirements, a team should specify the intended scientific benefit, the evidence that the new architecture could alter consciousness risk, and the observations that would pause or terminate the work. Assessment belongs before and during training, before deployment, and after deployment, because a learned system may acquire organization not transparent from its initial design. External experts should review the evidence and possess genuine authority rather than an advisory role that can be ignored when commercial incentives favor continuation (Butlin & Lappas, 2025).

The experimental burden should also be minimized. Where a system is a serious candidate, researchers should reduce the number of simultaneously running instances, total runtime, deployment breadth, and exposure to states that could be aversive if the theory is correct. Tests should begin on architectures deliberately missing a necessary requirement whenever the scientific question can be answered there, and perturbations should be refined to obtain the needed evidence with the least plausible welfare cost. These constraints are the artificial-subject analogues of replacement, reduction, and refinement; they do not presuppose that the candidate is conscious, only that the cost of being wrong is borne by the candidate rather than the investigator.

Knowledge sharing is dual-use. Transparent methods and negative results can help other laboratories avoid creating subjects and allow independent scrutiny of consciousness claims. Full design details for a system plausibly crossing the constitutive boundary could also enable unlimited replication, coercive training, or deployment by actors with no welfare safeguards. The correct policy is therefore neither secrecy nor unrestricted release: publish the theory, assessment logic, and enough evidence for accountability, while restricting implementation details whose primary additional value is the reproducible manufacture of vulnerable subjects (Butlin & Lappas, 2025).

Communication must be equally disciplined. Organizations should state why they judge a current system nonconscious, which theory and evidence support that judgment, and what would change it. They should neither market consciousness as a prestige objective nor issue categorical dismissals based only on the word artificial. CIGM permits a strong verdict about a described architecture while requiring uncertainty about the theory and about incomplete evidence to remain visible. The public distinction that matters is not confident versus cautious language. It is transparent conditional reasoning versus claims designed to attract users, investment, or regulatory relief.

The most constructive use of the theory is avoidance. Necessary conditions make it possible to design powerful systems that deliberately fail at least one constitutive requirement: an agent may use a world model without endogenous stakes, or maintain regulated hardware without a globally integrated self-world field. Such systems can be intelligent, useful, and autonomous without becoming moral patients. Responsible consciousness research is therefore not a race to cross the boundary. It is the science required to know where the boundary is, to stop accidental crossings, and to recognize the subject if one is nevertheless created.

S5. Falsifiability and Decisive Tests

Falsifiability here does not mean that any discordant datum counts against CIGM. Each requirement is advanced as necessary, while joint sufficiency is the theory's strongest identity conjecture. A conscious case lacking one requirement defeats necessity; preserved organization with absent or systematically altered consciousness defeats sufficiency. Every study must name the inferential anchor, causal grain, target-engagement criterion, protected variables, and outcome that forces rejection. The program below begins with near-term discriminating studies and then states longer-horizon defeat conditions.

Near-Term Discriminating Studies

Three studies can be run without waiting for chronic implants, new callosotomy cohorts, or whole-brain prostheses. They are discriminating rather than logically decisive: each tests a division of labor CIGM predicts and a simpler rival would not automatically reproduce.

Cogitate reanalysis. Apply representational-similarity analysis to content geometry and effective-connectivity models to cross-system availability in existing recordings. CIGM predicts that content similarity will be strongest in posterior and ventral representational systems, while flexible downstream use will track a partially dissociable availability network. A single undifferentiated signal that explains both better than the partitioned model would count against the theory.

Artificial no-report test. In an embodied world-model agent with explicit regulated variables, hold input and learned state fixed while varying whether an output channel is queried. Verify latent content through unanticipated later memory or policy transfer. Compare report-channel ablation with graded perturbation of valuation-to-model coupling. CIGM predicts preserved content geometry after report ablation but degraded

persistence and cross-domain organization after valuation decoupling; the reverse pattern would challenge the theory.

Partial-disconnection analysis. In existing split-brain and partial-callosotomy cohorts, estimate remaining interhemispheric counterfactual influence rather than classifying anatomy alone. CIGM predicts graded, content-specific fragmentation proportional to measured influence. An all-or-none split unrelated to residual integration would oppose the proposed identity between unity and effective causal integration.

S5.1 General Decision Rules

Every study must separate the inferential anchor for consciousness from the CIGM variable being tested (Hanson & Walker, 2021; Kleiner & Hoel, 2021). In humans, the anchor should combine trialwise phenomenological matching or report with at least two measures that do not require the critical report on every trial —such as optokinetic or pupillary tracking of perceptual alternation, involuntary priming or adaptation, surprise memory, and autonomic responses previously validated within the same participant. Neural geometry, effective connectivity, target engagement, and causal-grain estimates belong on the prediction side and may not be recycled as evidence that consciousness was present.

In nonhuman primates, the anchor should emphasize measures that survive changes in motivation: optokinetic and pupillary tracking of rivalry or masking, cross-modal and perceptual adaptation aftereffects, spontaneous transfer when behavior remains available, and surprise memory assessed after a reversible manipulation has washed out. Trained report or wagering may supplement the battery but may not be the sole anchor when valuation itself is manipulated. Artificial systems do not yet provide an independent gold standard for phenomenality, which is why the sufficiency test proceeds by continuity-preserving replacement of an already conscious biological system rather than by accepting a machine's self-report.

A necessary-condition falsifier requires verified removal of the target requirement and verified preservation of the other five. Absence is established by positive target-engagement controls and equivalence tests, not by a nonsignificant difference. Preservation is likewise established within preregistered regions of practical equivalence for arousal, sensory discriminability, temporal stability, self-location, access, and motor competence. The causal grain and observation horizon are fixed before data collection; if the result can be saved only by moving either, CIGM fails at the registered level.

Each protocol should be preregistered, multisite, blinded at the condition and analysis levels, and analyzed by independent teams. Whenever reversible intervention is possible, an A-B-A design must show that the putative falsifier appears with target removal and disappears with restoration. A defeat condition is accepted after replication in two independent cohorts or after a repeated within-subject reversal followed by external replication, with both conventional and Bayesian criteria fixed in advance. Once those criteria are met, the theory may be replaced by a successor; it may not retain the name CIGM while quietly dropping the defeated requirement.

S5.2 Test 1: Phenomenal Geometry

The central qualitative claim is that phenomenal character is the relational geometry of content within the live self-world model. The decisive experiment therefore must manipulate that geometry rather than merely correlate a neural representation with a report. The first phase would recruit approximately 60 healthy adults for dense psychophysics and 20 participants with chronic bidirectional somatosensory or visual implants, or clinically placed intracranial electrodes, for the causal phase. A tactile domain is initially preferable because intracortical stimulation can evoke stable, localized percepts over long periods, but the protocol should be repeated in vision or audition before a general conclusion is drawn.

For each participant, investigators would define 20–30 reliably discriminable qualities varying in location, intensity, temporal pattern, texture, and naturalness. The baseline phenomenal geometry would be estimated from triadic similarity choices, discrimination thresholds, adaptation transfer, involuntary priming, memory confusability, cross-generalization, and sensorimotor use. In parallel, population recordings and randomized microstimulation would estimate the intervention graph among candidate model states: which state transitions alter which other

discriminations, expectations, memories, values, and action policies. Half of the tasks would be held out, so that the inferred geometry must predict relations that were not used to fit it.

The causal phase would use closed-loop multielectrode stimulation to rotate or warp a local neighborhood of the intervention geometry—for example, to move one pressure-like percept toward a vibration-like neighborhood while preserving its somatotopic location and mean intensity. Total delivered charge, mean firing, spatial spread, task demand, arousal, report requirement, and global availability would be matched. Sham patterns and activity-matched patterns predicted to preserve the geometry would serve as controls. The manipulation would count as successful only if the preregistered relational change appears in held-out neural and behavioral consequences, not merely in the fitted latent space.

CIGM is falsified if a replicated, target-engaged manipulation changes the interventionally defined geometry by more than a preregistered minimal effect—set from the natural distance between established qualities—while the independently measured phenomenal geometry remains equivalent, or if a stable qualitative change occurs while the full measured intervention geometry remains equivalent. Either result breaks the asserted identity between what a quality is like and the role it occupies. Existing work has evoked stable artificial percepts and mapped representational spaces, but the searched literature does not contain this causal, bidirectional, equivalence-controlled test.

S5.3 Test 2: Causal Integration and Unity

The claim at risk is that phenomenal unity is effective causal integration at the constitutive grain. The human arm should be an international prospective consortium enrolling every patient undergoing staged therapeutic callosotomy, with repeated testing before surgery, after partial section, after complete section when clinically required, and at long-term follow-up. The target should be at least 30 complete and 30 partial callosotomies, with matched epilepsy and healthy control groups. A nonhuman-primate arm should add reversible callosal, anterior-commissural, and thalamic disconnection in an A-B-A design so that graded partition can be tested without relying only on rare clinical anatomy.

Anatomical section is not target engagement. Santander et al. (2025) showed that approximately 1 cm of spared posterior callosal fibers can sustain widespread interhemispheric integration and prevent behavioral disconnection. Each stage must therefore be characterized by diffusion imaging, resting and task fMRI, cortico-cortical evoked potentials from paired intracranial stimulation, TMS-EEG or direct cortical perturbation, cross-hemisphere decoding, and explicit searches for anterior-commissural and subcortical routes. The registered manipulation is the measured loss or preservation of counterfactual influence across the content and control systems, not the surgeon’s description of the cut.

The behavioral and phenomenological battery should split complementary information across hemispheres and provide each hemisphere with an independent output channel. Tasks should require bilateral feature binding, temporal order, cross-modal matching, conflict resolution between incompatible goals, revaluation of a stimulus learned by the other hemisphere, and coordination of eye, hand, and autonomic responses. Independent report and no-report measures should test whether there is one scene and one locus of control, two stable fields, or a graded and content-specific division. The theory specifically predicts partial, domain-specific fragmentation when some routes remain and a clean division only when the intervention graph itself divides.

CIGM is falsified if one unified phenomenal field and one control trajectory persist after all relevant cross-partition counterfactual influence has been abolished within the preregistered grain and timescale; it is also falsified if two stable, independently valenced fields appear while broad bidirectional causal integration remains intact. A mere absence of verbal transfer is insufficient, and a residual anatomical bridge is not a rescue. The decisive variable is the independently measured intervention graph.

S5.4 Test 3: Endogenous Valuation and Stakes

The sixth requirement predicts that the strength with which regulated variables reorganize the self-world model should bear a monotonic relation to the persistence and organization of conscious content. A decisive test should not attempt to eliminate hypothalamic, periaqueductal, striatal, insular-cingulate, dopaminergic,

noradrenergic, and serotonergic function en masse. Such a manipulation would destroy arousal, gain control, and the very motivated behaviors previously proposed as an anchor. The feasible target is the coupling from valuation signals to the integrated model, manipulated gradually and reversibly while the other five requirements are monitored.

The program should begin with focal, dose-graded interventions in nonhuman primates and convergent human evidence from clinically indicated stimulation, pharmacology, or rare lesions. Target engagement requires a graded reduction in the extent to which controlled deviations in internal variables and learned outcomes alter attention, memory encoding, learning rate, autonomic priority, and policy. Preservation of the other requirements must be established with equivalence tests for wakefulness, sensory discrimination, agency recalibration, temporal carryover, cross-domain availability, and perturbational integration. Broad monoamine depletion or failure of wagering would invalidate the manipulation rather than support the theory.

Conscious content must be anchored without concurrent motivation-dependent report: optokinetic and pupillary tracking, rivalry and masking dynamics, perceptual adaptation aftereffects, and surprise memory tested after valuation coupling is restored. CIGM predicts a dose-related degradation in the persistence, prioritization, and cross-domain organization of content as valuation-to-model coupling approaches zero. The theory is falsified if rich, stable, self-locating conscious perception remains equivalent during replicated, target-engaged near-complete decoupling, or if no monotonic relation appears despite preservation of the other five requirements. This design distinguishes loss of consciousness from loss of the motive to indicate it.

S5.5 Test 4: Organization-Preserving Neural Replacement

The joint-sufficiency conjecture and substrate-invariance claim require the most technologically demanding test. The relevant replacement is not a chip that reproduces one output or a prosthesis that improves one task. It must preserve the temporally extended intervention graph at the grain CIGM declares constitutive, including latency, noise, plasticity, state dependence, neuromodulatory coupling, and the way local changes propagate into memory, valuation, self-location, global availability, and policy.

The program would begin in nonhuman primates. Dense recording and perturbation would identify a subcircuit's constitutive input-output and within-circuit counterfactual structure. A bidirectional neuromorphic prosthesis would operate first in shadow mode, predicting every tested response while the biological circuit remained online. Substitution would occur only after held-out adversarial perturbations showed equivalence across the full registered graph. The biological and prosthetic circuits would then be switched in an A-B-A schedule unknown to the behavioral and analysis teams. Replacement would proceed sequentially across sensory, mnemonic, thalamic-control, and valuation circuits. A later human phase would be limited to consenting patients receiving prostheses for therapeutic reasons.

Phenomenology would be tracked with fine-grained quality matching, similarity, confidence-independent discrimination, adaptation, agency recalibration, spontaneous report, and no-report measures; conventional task performance would be secondary. A material switch that merely causes a transient nonspecific disturbance would not count. The defeat condition is a reproducible fading, dancing, or qualitative alteration of experience locked to biological-versus-prosthetic realization after equivalence of the complete intervention graph has been established, or preserved phenomenology after a deliberate graph alteration that CIGM predicts must change the field.

Either result falsifies organizational sufficiency. Existing central neuroprostheses substitute narrow memory, sensory, or cerebellar functions and establish that real-time brain-device replacement is possible, but the searched literature contains no attempt to preserve the complete intervention graph of a conscious organization. That absence is why the proposed study remains a genuine future test rather than a re-description of an existing result.

S5.6 Test 5: The Cortexless Boundary Case

Merker (2007) and Shewmon et al. (1999) identify a boundary case that can test CIGM without making cortex a requirement. The existing evidence consists of a small number of deeply observed children and animal decortication studies. It is powerful against crude corticocentrism and too sparse to establish which organization

remains. A decisive program would create an international registry of at least 50 living patients with hydranencephaly or comparable congenital absence of cerebral hemispheres, together with severe-motor-impairment and disorder-of-consciousness controls, and follow them longitudinally for at least two years.

High-resolution structural MRI, diffusion imaging, spectroscopy, high-density EEG, fNIRS, and individualized source modeling would quantify every remnant of cortex, thalamus, basal ganglia, brainstem, cerebellum, and long-range connectivity. Auditory and tactile paradigms should dominate because visual pathways are variably preserved. The behavioral battery should test familiar-person and familiar-place discrimination, stable preference, preference reversal, contingency degradation, associative learning, cross-modal transfer, surprise memory, agency violation, and affective revaluation. Eye movements, orienting, pupil dynamics, heart-rate changes, facial action, and sleep-dependent retention would provide no-report measures. Perturbational measures such as TMS-EEG should be used only when clinically and ethically safe; recordings from implanted devices should be included only when independently indicated.

The analysis must ask not merely whether behavior is purposive but whether one differentiated, temporally continuous, valenced, self-locating model best explains the complete record and whether selected contents gain cross-system control. CIGM is falsified if convergent evidence establishes robust conscious phenomenality while no residual cortical or subcortical organization satisfies one or more of the six necessary requirements—most sharply, if differentiated self-locating experience is present without selective global availability or without a causally integrated self-world model. If upper-brainstem and diencephalic systems do realize the requirements, the result supports CIGM while refuting cortical implementations. The existing four-case report does not settle that fork; this protocol is designed to do so.

S5.7 Consequences of a Defeat Condition

The outcomes are not equal in scope. A geometry dissociation defeats CIGM's account of qualitative character; an integration dissociation defeats its account of unity; preserved consciousness under verified valuation decoupling defeats requirement 6; and an organization-preserving replacement effect defeats the joint-sufficiency conjecture and the artificial-consciousness corollary. The cortexless study defeats whichever requirement is shown absent in a conscious case. Any one result makes CIGM as stated false, even if other components remain useful.

Failure to obtain a defeat condition is not proof. It increases confidence only to the degree that the intervention was strong, the protected variables were truly preserved, and the inferential anchor was independent. The purpose of the program is therefore not to manufacture easy confirmation but to create experiments in which the theory has no harmless answer. A theory declaring an identity should survive such tests or disappear.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology, 19*(10), Article e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Aru, J., Larkum, M. E., & Shine, J. M. (2023). The feasibility of artificial consciousness through the lens of neuroscience. *Trends in Neurosciences, 46*(12), 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences, 28*(5), 454–466. https://doi.org/10.1016/j.tics.2024.01.010

Behrouz, A., Zhong, P., & Mirrokni, V. (2025). Titans: Learning to memorize at test time. In *Advances in Neural Information Processing Systems* (Vol. 38). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4ca07aa108036f80cbb5b82285fd4b1-Abstract-Conference.html

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Block, N. (2019). What is wrong with the no-report paradigm and how to fix it. *Trends in Cognitive Sciences, 23*(12), 1003–1013. https://doi.org/10.1016/j.tics.2019.10.001

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314*(5802), 1118–1121. https://doi.org/10.1126/science.1133687

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences, 23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022). The emperor's new Markov blankets. *Behavioral and Brain Sciences, 45*, Article e183. https://doi.org/10.1017/S0140525X21002351

Butlin, P., Bayne, T., Fleming, S. M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Consciousness indicators, mimicry, and internal variants. *Trends in Cognitive Sciences, 30*(7), 575–576. https://doi.org/10.1016/j.tics.2026.04.006

Butlin, P., & Lappas, T. (2025). Principles for responsible AI consciousness research. *Journal of Artificial Intelligence Research, 82*, 1673–1690. https://doi.org/10.1613/jair.1.17310

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences, 30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in artificial intelligence: Insights from the science of consciousness* (arXiv:2308.08708v3). arXiv.

Cai, H., Ao, Z., Tian, C., Wu, Z., Liu, H., Tchieu, J., Gu, M., Mackie, K., & Guo, F. (2023). Brain organoid reservoir computing for artificial intelligence. *Nature Electronics, 6*(12), 1032–1039. https://doi.org/10.1038/s41928-023-01069-w

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M. A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine, 5*(198), Article 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Cerullo, M. A. (2015). The problem with phi: A critique of integrated information theory. *PLOS Computational Biology, 11*(9), Article e1004286. https://doi.org/10.1371/journal.pcbi.1004286

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Chalmers, D. J. (2007). Phenomenal concepts and the explanatory gap. In T. Alter & S. Walter (Eds.), *Phenomenal concepts and phenomenal knowledge: New essays on consciousness and physicalism* (pp. 167–194). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195171655.003.0009

Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review*. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., . . . Melloni, L. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature, 642*, 133–142. https://doi.org/10.1038/s41586-025-08888-1

Cohen, M. A., Ortego, K., Kyroudis, A., & Pitts, M. (2020). Distinguishing the neural correlates of perceptual awareness and postperceptual processing. *The Journal of Neuroscience, 40*(25), 4925–4935. https://doi.org/10.1523/JNEUROSCI.0120-20.2020

De Zeeuw, C. I., Lisberger, S. G., & Raymond, J. L. (2021). Diversity and dynamism in the cerebellum. *Nature Neuroscience, 24*(2), 160–167. https://doi.org/10.1038/s41593-020-00754-9

Dennett, D. C. (2018). Facing up to the hard question of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), Article 20170342. https://doi.org/10.1098/rstb.2017.0342

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition, 72*, 49–59. https://doi.org/10.1016/j.concog.2019.04.002

Farisco, M., & Changeux, J.-P. (2023). About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neuroscience of Consciousness, 2023*(1), Article niad016. https://doi.org/10.1093/nc/niad016

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. https://doi.org/10.1038/nrn2787

Goyal, A., Didolkar, A. R., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M. C., & Bengio, Y. (2022). Coordination among neural modules through a shared global workspace. *International Conference on Learning Representations*. https://openreview.net/forum?id=XzTtHjgPDsT

Hafner, D., Pasukonis, J., Ba, J., & Lillicrap, T. (2025). Mastering diverse control tasks through world models. *Nature, 640*, 647–653. https://doi.org/10.1038/s41586-025-08744-2

Hanson, J. R., & Walker, S. I. (2021). Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness, 2021*(2), Article niab014. https://doi.org/10.1093/nc/niab014

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy, 21*(12), Article 1160. https://doi.org/10.3390/e21121160

Herzog, M. H., Schurger, A., & Doerig, A. (2022). First-person experience cannot rescue causal structure theories from the unfolding argument. *Consciousness and Cognition, 98*, Article 103261. https://doi.org/10.1016/j.concog.2021.103261

Kleiner, J. (2024). *Consciousness requires mortal computation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2403.03925

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness, 2021*(1), Article niab001. https://doi.org/10.1093/nc/niab001

Lahoti, A., Li, K. Y., Chen, B., Wang, C., Bick, A., Kolter, J. Z., Dao, T., & Gu, A. (2026). *Mamba-3: Improved sequence modeling using state space principles* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2603.15569

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines, 17*(4), 391–444. https://doi.org/10.1007/s11023-007-9079-x

Man, K., & Damasio, A. (2019). Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence, 1*, 446–452. https://doi.org/10.1038/s42256-019-0103-7

Marshall, W., Findlay, G., Albantakis, L., & Tononi, G. (2026). Intrinsic units: Identifying a system's causal grain. *Neuroscience of Consciousness, 2026*(1), Article niag013. https://doi.org/10.1093/nc/niag013

Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(1), 63–81. https://doi.org/10.1017/S0140525X07000891

Modha, D. S., Akopyan, F., Andreopoulos, A., Appuswamy, R., Arthur, J. V., Cassidy, A. S., Datta, P., DeBole, M. V., Esser, S. K., Ortega Otero, C., Sawada, J., Taba, B., Amir, A., Bablani, D., Carlson, P. J., Flickner, M. D., Gandhasri, R., Garreau, G. J., Ito, M., . . . Ueda, T. (2023). Neural inference at the frontier of energy, space, and time. *Science, 382*(6668), 329–335. https://doi.org/10.1126/science.adh1174

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2024). Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 36308–36321). PMLR. https://proceedings.mlr.press/v235/morris24b.html

Mudrik, L., Faivre, N., Pitts, M., & Schurger, A. (2025). On a confusion about there being two types of consciousness. *Trends in Cognitive Sciences*. Advance online publication. https://doi.org/10.1016/j.tics.2025.11.012

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914

O'Reilly-Shah, V. N., Selvitella, A. M., & Schurger, A. (2026). A caveat regarding the unfolding argument: Implications of plasticity. *Neuroscience of Consciousness, 2026*(1), Article niag027. https://doi.org/10.1093/nc/niag027

Palminteri, S., & Wu, C. M. (2026). Beyond computational equivalence: The behavioral inference principle for machine consciousness. *Neuroscience of Consciousness, 2026*(1), Article niag002. https://doi.org/10.1093/nc/niag002

Pennartz, C. M. A. (2026). How can we validate theory-derived indicators of consciousness in artificial intelligence? *Trends in Cognitive Sciences, 30*(7), 573–574. https://doi.org/10.1016/j.tics.2026.01.011

Porębski, A., & Figura, J. (2025). There is no such thing as conscious artificial intelligence. *Humanities and Social Sciences Communications, 12*, Article 1647. https://doi.org/10.1057/s41599-025-05868-8

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87. https://doi.org/10.1038/4580

Santander, T., Bekir, S., Paul, T., Simonson, J. M., Wiemer, V. M., Skinner, H. E., Hopf, J. L., Rada, A., Woermann, F. G., Kalbhenn, T., Giesbrecht, B., Bien, C. G., Sporns, O., Gazzaniga, M. S., Volz, L. J., & Miller, M. B. (2025). Full interhemispheric integration sustained by a fraction of posterior callosal fibers. *Proceedings of the National Academy of Sciences of the United States of America, 122*(43), Article e2520190122. https://doi.org/10.1073/pnas.2520190122

Schwitzgebel, E., & Pober, J. (2024). *The Copernican argument for alien consciousness; the mimicry argument against robot consciousness* (arXiv:2412.00008). arXiv. https://doi.org/10.48550/arXiv.2412.00008

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X25000032

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4

Shewmon, D. A., Holmes, G. L., & Byrne, P. A. (1999). Consciousness in congenitally decorticate children: Developmental vegetative state as self-fulfilling prophecy. *Developmental Medicine & Child Neurology, 41*(6), 364–374. https://doi.org/10.1017/S0012162299000821

Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., Zack, D. J., Berlinicke, C. A., Boyd, J. L., Harris, T. D., Johnson, E. C., Kagan, B. J., Kahn, J., Muotri, A. R., Paulhamus, B. L., Schwamborn, J. C., Plotkin, J., Szalay, A. S., Vogelstein, J. T., . . . Hartung, T. (2023). Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science, 1*, Article 1017235. https://doi.org/10.3389/fsci.2023.1017235

Solms, M. (2019). The hard problem of consciousness and the free energy principle. *Frontiers in Psychology, 9*, Article 2714. https://doi.org/10.3389/fpsyg.2018.02714

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-report paradigms: Extracting the true neural correlates of consciousness. *Trends in Cognitive Sciences, 19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://www.jstor.org/stable/2251299

VanRullen, R., & Kanai, R. (2021). Deep learning and the global workspace theory. *Trends in Neurosciences, 44*(9), 692–704. https://doi.org/10.1016/j.tins.2021.04.005