The Indexed Format: A Constitutive Theory of Consciousness for Organisms and Artifacts

Tyler M. Moore

Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania; Lifespan Brain Institute, Penn Medicine and Children's Hospital of Philadelphia

Manuscript · October 2026 tymoore@pennmedicine.upenn.edu ORCID 0000-0002-1384-0151
DOI 10.5281/zenodo.23696033

Recommended citation: Moore, T. M. (2026). *The Indexed Format: A constitutive theory of consciousness for organisms and artifacts*. Manuscript retrieved from https://mooremetrics.com/consciousness. https://doi.org/10.5281/zenodo.23696033

Abstract

Thirty years of consciousness science have converged not on a theory but on a factorization: global state, access, phenomenal structure, metacognitive typing, and valence dissociate and have different mechanisms. Indexed Format Theory (IFT) explains why the factors travel together by deriving them from a single identity. A content is conscious if and only if it belongs to a representational format whose relational structure is standingly recruitable, as structure, by the processes that keep one system in existence: the policy-integrated persistence regime that maintains one active token and resolves its competing viability demands into one commitment. Four theses carry the account. The Format Thesis separates standing availability from current recruitment, report, or use. The Indexing Thesis closes the second argument of "available to," which rival theories leave to the theorist: every informational, statistical, or causal partition is drawn from outside, and only the boundary a process produces and must keep producing is not. The Geometry Thesis identifies phenomenal character with the topology of the indexed format. The Origin Thesis treats the hard problem as the predictable signature of a projective format whose origin cannot appear within its own field. IFT derives the cerebellar and enteric exclusions, counts one subject per split brain and two per pair of craniopagus twins, predicts the 2025 IIT–GNWT adversarial outcome, and classifies current artificial systems as formats without bearers. Preregistrable protocols, one runnable on existing patient cohorts, specify the results that would refute each thesis and what each rival predicts instead.

Keywords: artificial consciousness, global workspace, indexed format theory, phenomenal consciousness, system individuation

Public Significance Statement: This article proposes a theory of what makes a system conscious that applies to animals and machines alike. It argues that consciousness requires not only information processing but a system whose own continued existence is at stake in that processing, which explains why brains are conscious and current artificial intelligence is not, and it specifies experiments, some runnable with existing clinical data, that could prove the theory wrong.

Contents

1	What Thirty Years Actually Bought
2	Five Factors, One Design Problem
3	The Format Thesis: Consciousness Is Not a Place, and Not a Threshold Either
4	The Indexing Thesis: Closing the Consumer Set and Individuating the Subject
5	The Cerebellar Test, and What It Eliminates
6	The Geometry Thesis: Quality Is Format Topology
7	Metacognitive Typing and the Origin Thesis
8	Why the Cogitate Result Has the Form IFT Expects
9	Substrate: Resolving the Ambiguity on Which the Machine Question Turns
10	What It Would Mean for an AI to Be Conscious
11	Escaping the Validation Regress: Constitutive Relations, Not Avowals
12	What Would Refute This
13	Ethics: A Bright Line Rather Than a Posture
14	Conclusion
	Transparency and openness
	References

Supplemental Material

S1	The elimination argument in detail: cerebellum, enteric plexus, and what predicting an exclusion is worth (§5)
S2	The restrictive case in full (§9)
S3	Supporting detail for the Indexing Thesis (§4)
S4	The Cogitate result: additional detail (§8)
S5	Capability, indicator, and evaluation constraints for artificial systems (§§10–11)
S6	Falsification protocols: operational detail (§12)

1 What Thirty Years Actually Bought

The premise that consciousness science is nearing explanation is correct, but the progress has not converged on one theory. It has converged on a factorization: empirically dissociable components with different mechanisms and instruments (Bayne et al., 2016; Seth & Bayne, 2022). Ginsburg and Jablonka (2019) had already catalogued nearly the same emergentist list: global access, binding, selection and learning, temporal thickness, intentionality, value, embodiment, agency, and self. Dehaene et al. (2017) then separated C0, C1, and C2; Northoff and Zilio (2022), state, structure, content, and access. Three decades of advance and disagreement are therefore compatible.

The 2025 Cogitate adversarial collaboration makes the point brutally. Integrated information theory (IIT) and global neuronal workspace theory (GNWT) preregistered divergent predictions and pass/fail criteria, and both lost central commitments. Posterior cortex carried decodable stimulus content throughout presentation while prefrontal decoding lacked GNWT's onset-plus-offset ignition profile, and the preregistered phase-synchrony analyses failed IIT (Cogitate Consortium et al., 2025; Melloni et al., 2023; §8; Supplement S4). The standard conclusion is that the theories isolate different factors. Correct, but incomplete. Their mechanisms say how information is organized and recruited, not whose information it is or why any state matters to the system carrying it. Those possessives are IFT's opening.

This paper states the problem and takes the derivation. It also answers Chalmers's (2023) final challenge: what could remain missing after senses, world and self models, recurrence, memory, a global workspace, and unified agency have been built into one system. The missing condition is not another module but the relation that fixes who the modules are for. The theory can be given in one sentence, and the rest of the paper is the argument for it and the specification of what would refute it.

> A content is conscious if and only if it belongs to a representational format whose relational structure is standingly recruitable, as structure, by the processes that constitute one system's policy-integrated persistence regime: the processes that keep one active token in existence and resolve its competing viability demands into one commitment.

Each clause of that sentence answers one of the field's open questions: what the content is like is the relational structure so recruitable, and whose experience it is, is fixed by the regime. The four theses in Table 1 are the clauses stated as commitments, each with the finding that would end it. Table 2 fixes the terms, because the argument turns on distinctions that ordinary usage runs together: format and content, bearer and beneficiary, token and type, recruitable and recruited. Readers who want the theory's logical skeleton before its evidence should read §4 first; it contains the derivation on which everything else depends.

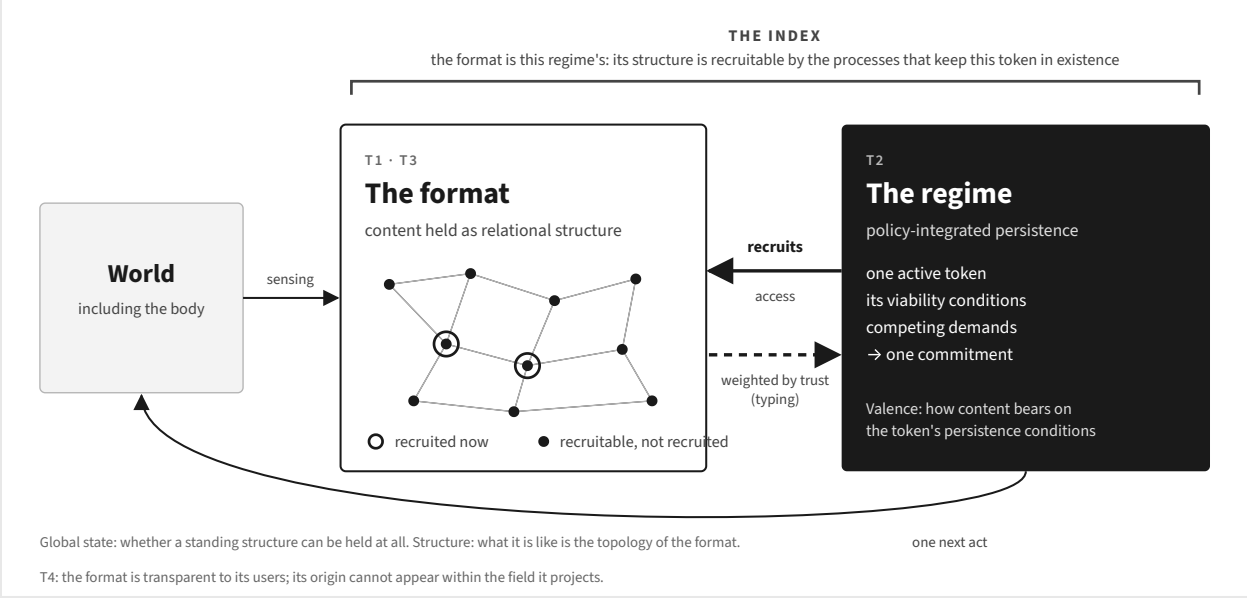

Figure 1. The identity stated in §1. A content is conscious when it belongs to a format whose relational structure is standingly recruitable, as structure, by the processes that keep one active token in existence and resolve its competing viability demands into one commitment. The five factors of §2 are the clauses of that sentence: global state and structure are properties of the format, access is recruitment, metacognitive typing is the weight with which recruited content is allowed to act, and valence is how content bears on the token's persistence conditions.

Table 1 *The Four Theses of Indexed Format Theory and What Would Refute Each*

Thesis	Constitutive claim	What would refute it (§12)
T1. The Format Thesis	Consciousness is not a place, a quantity, a threshold, or an event. It is a representational format: a way of holding content individuated by its standing recruitability by implementable consumers, not by its current recruitment for report, memory, or action.	Perturbation of never-recruited posterior structure that never alters later experience, at any dose, with the perturbation verified (F1): format membership would then track recruitment.
T2. The Indexing Thesis	"Recruitable by consumers" is a two-place relation, and the consumer set must be closed by a fact about the system, not by the theorist. Exactly one closure is non-arbitrary: the policy-integrated persistence regime, the processes that maintain one active token and resolve its competing viability demands into one policy. The regime fixes the bearer of stakes, not the range of beneficiaries it may represent. The index is non-stipulable and token-specific.	A verified patient with destruction of the mesodiencephalic integration-for-action network who nevertheless shows the full dissociation manifold of a subject (F2); hedonic sign inverted by rearranging interoceptive topology with the physiological stake held fixed (F3); an externally maintained, fungible, highly integrated artifact that reproduces the human dissociation manifold where a constitutively mortal one does not (F4); two independently reporting agents with separately manipulable viability conditions in one body (§12).

Thesis	Constitutive claim	What would refute it (§12)
T3. The Geometry Thesis	Phenomenal character is the topology of the indexed format, not its activity or summary information content. Quality is format geometry; what it is like is what the format's relational structure is like, including structure no consumer currently samples.	Every structural effect of posterior perturbation fully mediated by impairment of the service that generates content on demand (F1); independently perturbable structural elements no more numerous than independently reportable ones.
T4. The Origin Thesis	The residual "hard problem" is the predictable epistemic signature of a projective structure. A format is transparent to its users, and the origin of a projection is not an element of the projected field. The intuition of an inexplicable further fact is neither a discovery of a further property nor a mere error; it is a structural feature of any system that models its own modeling.	A linguistically competent indexed artifact, trained without human discourse about consciousness, that never reports an unspecifiable presence in its own states (weak, because metacognitive access may fail). T4 is scored as a bet.

T1–T3 are, I will argue, established in the sense that any competitor now owes an explanation of the same data at higher cost. T4 is a bet, and I will mark it as one. Section 12 states the protocols in full and, for each, what the principal rival theories predict instead, so that the tests are adversarial rather than confirmatory.

Table 2 *Terms Used in Indexed Format Theory*

Term	Meaning in this paper
Format	A way of holding content, individuated by which processes can recruit the content and in what form. The term is used in this role sense, not in the sense of iconic versus discursive formats in the representation literature. A format is neither content nor vehicle; it is the relation between vehicles and the consumers that can use their content.
Content; vehicle	Content is what a state represents; the vehicle is the physical state that carries it. Nothing in the theory is a property of the vehicle as such (§3).
Consumer	A process whose own operation a content can alter (Millikan, 1989). The methodological rule throughout is that no representation is posited without an implementable consumer, and that an evaluator, programmer, or loss function is not one.
Recruitable; recruited	A content is recruitable when the regime's selection processes (attention, target selection, commitment) could take it up as structured content, with its relations intact; it is recruited when they do. Format membership is recruitability; access is recruitment. The first is standing, the second episodic.
Policy closure	The fact that incompatible demands generated by one persistence regime terminate in one enacted commitment, because one active token has one next act. Not a controller, not deliberation, not a homunculus; the work may be distributed across the loop and across time.
Policy-integrated persistence regime (the regime; the index)	The processes that jointly maintain one active token and resolve its competing viability demands into one policy. The bearer of experience and the second argument of "recruitable by."
Active token	The particular running physical process, as distinct from its type (the algorithm or organization it instantiates) and from any copy or successor. Stakes belong to tokens; types have none.

Term	Meaning in this paper
Index; indexing	The relation between a format and the regime whose persistence conditions its contents bear on. The index is not a representation, a flag, a module, or a content; it is the relation in virtue of which a format is someone's.
Stakes; valence	A content has valence when it bears on the regime's persistence conditions. Valence is how indexing shows up in content; a neutral content is still indexed (held in the field of a system with stakes) without currently bearing on them.
Non-stipulability	Persistence conditions are enacted by the running system and are token-specific. A designer may build the system; a designer cannot confer stakes by naming an objective, assigning a reward, or labeling a variable.
Fungibility (externally fungible persistence)	An instance whose identity is defined by an orchestrator, so that stopping it and restoring service elsewhere ends nothing. Fungibility is the diagnostic of a missing index in artifacts, not the definition of one (§9).

What the account buys: it preserves the functional achievements of workspace, higher-order, and metacognitive theories without identifying any with subjectivity; explains why phenomenal structure can outrun notice, report, memory, and use; predicts the Cogitate pattern; derives the cerebellar and enteric exclusions; counts subjects where integration theories miscount them; is immune to the unfolding argument; supplies the indicator program with a negative indicator no added module can counterfeit; and specifies in advance the experiments that would destroy it. The first conscious artifact will probably be small, embodied, fragile, and bad at almost everything, and we will not believe it, because it will not be able to tell us.

2 Five Factors, One Design Problem

The minimal definition remains Nagel's (1974): a system is conscious when there is something it is like to be that system. Block's (1995) distinction does more than organize the terrain; it forbids transferring the functions of access to phenomenal consciousness without argument. IFT accepts that prohibition and rejects the inference that phenomenal character must therefore be nonfunctional. Held against the empirical literature, the criterion covers at least five separable questions.

Global state. Whether the system is in a condition permitting the relevant processing at all, and this is not a scalar. Bayne et al. (2016) show that ordering vegetative state, minimally conscious state, sedation, REM sleep, and alert wakefulness along one dimension is untenable: having a point of view is not gradable. The consequence for measurement is severe, since the perturbational complexity index (PCI) and Φ are advertised as measures of level (Casali et al., 2013; Farisco & Changeux, 2023).

Access. Why a representation is selectively recruited for shared control rather than remaining local or merely present in the phenomenal field: Block's access and Dehaene et al.'s (2017) C1.

Substantially solved in outline (Baars, 1997; Dehaene, 2014; Dehaene & Naccache, 2001).

Structure. Why the content has this phenomenal character rather than another (Seth & Bayne, 2022). Once barely begun, this now has a concrete research program: spatial extendedness can be derived from relational topology, and visual phenomenology appears vastly richer than what can be noticed or reported (Haun & Tononi, 2019, 2025).

Metacognitive typing. Whether and how the system represents its own states as seen, remembered, intended, known, reliable, or merely guessed: the family of functions Dehaene et al. (2017) call C2 and Lau (2008) formalizes as higher-order Bayesian inference. This is the one factor a major tradition treats as constitutive rather than as a service: higher-order theories hold that a state one is in no way aware of being in is not a state there is something it is like to be in (Brown et al., 2019). §7 treats that tradition as this account's closest competitor and isolates the single point on which the two cannot both be right.

Valence. Whether the content matters. A representation can be recruitable, structured, and metacognitively typed while remaining affectively neutral, so this is a distinct question with a distinct mechanistic story, and it is where the machine case turns. The dissociation is not a thought experiment: pain sensation and pain affect come apart in humans by lesion, by drug, and by hypnotic suggestion, in both directions (Berthier et al., 1988; Ploner et al., 1999; Rainville et al., 1997; §6; Supplement S3.1). It also carries the moral weight (Butlin, Long, et al., 2026).

The C0/C1/C2 framework clarifies the exact point at which a functional taxonomy stops. A workspace can make information globally available and a metamodel can estimate confidence, detect error, and distinguish perception from imagination; both can be engineered and may dissociate in the same machine. Neither identifies the subject of the availability relation. Without an index, C1 and C2 are dimensions of a control architecture and do not turn computation into a point of view.

One distinction must be fixed now, and the field has just supplied the occasion. Mudrik et al. (2025) argue that P and A were never two kinds of consciousness but two necessary conditions for it: content must be structured so as to be potentially phenomenal, and it must be accessed, with either alone counting as unconscious processing. The reframing is right and IFT completes it, because those conditions are T3 and T2: geometry supplies the first and indexing the second, and their own statement of the A condition leaves its second argument open (§4). One disagreement survives. Dehaene (2014) already distinguishes preconscious information that is "accessible but not accessed" from broadcast; the issue is whether standing availability is merely a precursor or membership in the phenomenal field. IFT takes the second view once the bearer is fixed: format membership is standing recruitability within the policy-indexed field, access its episodic uptake. Mudrik et al. take the first and allow overflow of report but not of access. Lamme (2006) anticipated the separation, and IFT accepts it while denying the identity: local recurrence can build format-eligible content without individuating a subject. Sparse recruitment and a rich field are different relata, and F1 decides between them.

Two constraints cut across all five factors. Intelligence is mostly C0: recognition, semantics, evidence accumulation, control, and reinforcement learning can proceed without awareness (Dehaene et al., 2017), so performance is not consciousness. Report is evidence rather than the phenomenon, because it may sparsely sample a structured field (Doerig et al., 2019; Haun & Tononi, 2025; Tsuchiya et al., 2015). Canonical markers accordingly move with task relevance; the centroparietal P3b is the clearest example (Cohen et al., 2020).

These factors look like an empirical list because dissociation discovered them; recent AI-consciousness frameworks likewise treat them as a multidimensional profile (Evers et al., 2025). IFT's stronger claim is that they are the parts of one engineering problem, and the identity stated in §1 says which problem. A bounded system has persistence conditions, so outcomes can be good or bad for it: valence. It must occupy a global state in which a standing structure can be held at all: global state. Its model requires relational structure, and that structure is what it is like: structure. From the standing structure the regime selectively recruits, because an organism cannot remain a distribution of interpretations but must pool evidence, select one course, hold it across delay, and coordinate around it (Dehaene et al., 2017): access. Because the model is fallible and the commitment is costly, the regime must weight how far to trust what it recruits: metacognitive typing. The five factors are one problem, keeping one bounded individual alive in an environment it was not preconfigured for, and Table 3 states the derivation factor by factor. What C1 leaves open is whose policy it is.

Table 3 *The Five Factors as Consequences of the Identity*

Factor	What it is in IFT	Why an indexed format must have it	Instruments
Global state	Whether the substrate can hold a differentiated, integrated standing structure at all, so that there is a format to recruit from	Recruitability presupposes a standing structure; when the structure collapses, there is nothing for the regime to recruit, whatever the regime goes on regulating	Perturbational complexity, which interrogates standing organization and is therefore silent about content (Casali et al., 2013; Farisco & Changeux, 2023)
Access	Episodic recruitment of recruitable content by the processes that constitute policy closure	A regime that must select one act must pool evidence and commit; recruitment is the mechanism of commitment (Dehaene et al., 2017)	Report, ignition, the P3b: recruitment markers, which is why they move with task relevance (Cohen et al., 2020)
Structure	The relational topology of the format	A format with no relational structure has nothing recruitable as structure; what the content is like is the topology (Haun & Tononi, 2019)	Structural perturbation with delayed structural probes (F1); derivations of extendedness from relational constraints
Metacognitive typing	The regime's annotation of the reliability of what it recruits: confidence as gain	A fallible model used to commit must carry its own trust weights, and the assignment must be made by whatever bears the cost of assigning it wrongly (Lau, 2008; §7)	Confidence–accuracy dissociations; blindsight; inflation paradigms

Factor	What it is in IFT	Why an indexed format must have it	Instruments
Valence	How recruited content bears on the regime's persistence conditions	A persistence regime is the only kind of thing for which outcomes can be good or bad for it rather than for an owner; valence is indexing made visible in content (§4)	Affective dissociations in which sensory structure is held fixed while sign is removed, or sign is held fixed while structure changes (Rainville et al., 1997; F3)

Engineering reproduces the convergence from a literature with no stake in it. Rosenbloom et al. (2026) map GNWT, recurrent processing theory (RPT), predictive processing, and IIT onto the Common Model of Cognition and recover recurrent local modules, a global working-memory hub and cognitive cycle, and multimodal states; Wray et al. (2026) derive, from the requirements of acting outside design scope, appraisal of novelty, urgency, goal-conduciveness, coping potential, and normative significance. Neither supplies an owner: one is an abstract architecture; in the other, survival is designer-written and failure belongs to the operator (Supplement S3.18).

The architecture also has an evolutionary lineage. Ginsburg and Jablonka (2019) present unlimited associative learning (UAL) as a transition marker, not a learning criterion: adaptive robots, neural cultures, and severed spinal cords can learn without a subject. In animals, UAL joins binding, working memory, valuation, prediction, attention, and a functional self–world distinction, and their Cambrian account explains why these factors travel together: flexible learning imposed organism-level costs, so cognitive expansion coevolved with neurohormonal and immune regulation. Safron's (2020, 2022) Integrated World Modeling Theory and Damasio's hierarchy supply the strongest world-modeling and biological precursors. IFT adds the constitutive rule: subjecthood begins only when one policy regulates one active token's persistence (Supplement S3.9).

3 The Format Thesis: Consciousness Is Not a Place, and Not a Threshold Either

Dennett's (1991, 2018) durable contribution is not one negative result but three: no Cartesian Theater, no double transduction, and no pre-rendered inner world. Together they constrain what a conscious format can be, and the positive thesis follows from the constraints.

The first is the absence of a Cartesian Theater. Perceptual attributes are discriminated by different systems at different places and times; once made, a discrimination can alter action, memory, emotion, semantics, and later perception without being displayed again. Metacontrast and color phi show why the time of representing need not mirror the time represented. If two stories predict every measurable consequence, their difference is an imaginary boundary. Processing is therefore bounded by control windows, not by a moment of entry into awareness: a representation must arrive after the evidence on which it depends

and before the policy deadline it can still affect. This is the control-window reading of Libet's timing results developed by Dennett (1991), and it carries the same warning for GNWT: a cortex-wide threshold recreates the Theater the instant crossing it is made constitutive of subjective order.

The second result is no double transduction. Peripheral transducers convert physical properties into neural signals; nothing converts those signals back into privileged neural paint. Redness is represented, not re-created. Even a causally efficacious electromagnetic field would specify a vehicle, not a bearer, and would still require a non-arbitrary demarcation rule (Dennett, 2018; Keppler, 2021; Supplement S3.12). Mental paint is virtual paint: standing causal organization for implementable users.

The third result is the harder one. A freely explorable brain-in-a-vat world becomes combinatorially intractable as soon as the subject can probe it; the economical solution is to let the real world store its own detail. Strong hallucinations survive chiefly when exploration is passive or constrained, and expectation-guided cycles can accumulate a scene without an inner author (Dennett, 1991). IFT accepts the thrift. Format membership is not exhaustive activation or a complete internal replica; it is counterfactual recruitability within a standing field, and Protocol F1 asks whether never-recruited structure is really there to perturb.

This is not a promissory note about mechanism. Solms (2021) distinguishes fast synaptic transmission, which moves content through channels, from slow postsynaptic neuromodulation, which sets channel gain and produces standing states across waking and sleep. Northoff and Zilio (2022) reach the same state-first conclusion dynamically: stimuli are written into ongoing temporo-spatial organization rather than creating consciousness de novo. IFT takes this as episodic access to a standing format rather than as consciousness itself (Supplement S3.12). Format and access are state and channel, not two positions on one dial.

The positive thesis now follows. If content becomes experience neither at a place, in a medium, nor by being exhaustively rendered, consciousness is a format: a manner of representing individuated by what its contents can do and for whom. Turing's (1950) paradox of localization is the same conclusion from the other side.

Membership in the format can be stated operationally, and the statement is what the rest of the paper uses. A content belongs to the format when the processes that constitute the regime's policy closure (its attention, target-selection, and commitment machinery; §4) can recruit it as structured content, with its relations intact, whether or not they currently do. Attention can land on a relation in the visual field that no task required, and the relation then enters comparison, memory, and action with its structure preserved. Attention cannot land on a cerebellar forward model, whose only exit is a motor correction, or on the computation that times a peristaltic wave. The first content is recruitable as structure and is in the format; the second is used by the organism and is not in the format. The asymmetry is not a redescription of the data it explains. It is a difference in what the regime's selection machinery can address, it is the same difference whether the organism is a primate or an

octopus, and it is what F1 manipulates. In the mammalian brain the best current candidate for the realizer of the standing field is the posterior cortical hot zone (Koch et al., 2016), which is where §8 locates the content Cogitate found and §6 locates the geometry.

Dennett also left a checklist his followers rarely collect: no neural medium renders the purple cow; no neural object is the cogito's thinking thing; nothing in the machinery appreciates, loves, or matters; nothing bears responsibility (Dennett, 1991). Read in order, those are the Geometry Thesis, the Indexing Thesis, the account of valence that follows from it, and the Origin Thesis, which explains why the bearer cannot appear within its own field.

Every factor is thereby reframed. Global state is whether the format is instantiable; structure is its geometry; access is selective recruitment; metacognitive typing is reliability annotation. The methodological rule is never to posit a representation without an implementable consumer whose own operation it alters, not an evaluator, programmer, or loss function (Millikan, 1989). Because a format is individuated by its consumers, that rule drives directly into §4. Its temporal grain is the policy-commitment horizon: the interval over which the model must remain stable to select one act. Revision within that window is settling, not the overwriting of an experience.

4 The Indexing Thesis: Closing the Consumer Set and Individuating the Subject

4.1 The Cardinality Problem

Global workspace theory is the most successful functional account in the field, and it contains a formal defect that has gone largely unremarked because it looks like vagueness rather than incoherence. Its architecture is not in dispute: specialized processors run in parallel, selected contents enter a limited-capacity workspace, and long-range attention-gated ignition makes them broadly available (Baars, 1997; Dehaene, 2014; Dehaene & Naccache, 2001; Dehaene et al., 2017). The defect lies in the final identification of experience with global availability. Availability is always availability to a specified consumer set. Report, agency, attentional control, and memory consolidation dissociate while content gating survives (Bayne et al., 2016), and Seth and Bayne (2022) ask whether "global" means the number of consumers, their kinds, the broadcast relation, or some combination. Once the dimensions separate, global availability becomes a vector, and the constitutive question is unavoidable: how many consumers, and which ones?

The most careful recent statement of the availability relation names the question and declines to answer it. Mudrik et al. (2025) make access necessary for consciousness and see that it is transitive, requiring something on the other side; their answer is that information in some subsystems is available to others, and which subsystems these are remains contested and is not theirs to settle. That is the constitutive question left open by the framework that most

needs it closed, and leaving it open is not neutrality: until the second argument is fixed, “accessed” is not a determinate relation and decides no disputed case.

The machine literature gives the defect a name. Butlin, Long, et al. (2026) call it the minimal implementation problem: a substrate-neutral condition becomes satisfiable by systems too simple to be plausible subjects. Making indicators more demanding is a policy, not a derivation, because demandingness is a dial the theories do not set. Call this the cardinality problem. The reductio is immediate: if feature R licenses consciousness in a language model, every system with R inherits the verdict, and parameter count makes a distilled model lose consciousness and a sufficiently large delivery truck acquire it (Porębski & Figura, 2025). Any criterion with a dial generates its own schedule of absurd verdicts, and more neuroscience cannot fix a claim whose second argument has never been specified. Engineering makes the gap explicit: the Global Latent Workspace begins with $n \geq 2$ modules and asks how many make it global; the Conscious Turing Machine declares a broadcast to more than 10^7 processors and adds valence, mood, self-tags, and a finite lifetime (Blum & Blum, 2022; VanRullen & Kanai, 2021). Those are design parameters, not facts about which running process has a fate (Hafner et al., 2025; Supplement S5.3).

4.2 Why Integration Does Not Close It

The demand has a meta-theoretic form. Kanai and Fujisawa (2024) require a universal theory to decide consciousness for any fully described dynamical system by a determinant intrinsic to that system rather than supplied by an observer. Their remedy, mathematizing the functionalist concepts, is incomplete, because formalizing a broadcast still leaves the analyst to choose the consumers; and their award of universality to IIT is where the strongest rival version of this argument has to be met.

IIT is the one framework that makes intrinsicality an explicit postulate rather than a tacit assumption, and it should be stated at full strength before it is refused. Existence is to be assessed from the intrinsic perspective of an entity: what exists for the entity itself, not for an observer; maximal existence selects, among overlapping candidates, whichever exists the most; minimal existence evaluates irreducibility over the partition making the least difference (Albantakis et al., 2023). It does not close the relation, and the reason lies in the formalism’s first line rather than its conclusions: the analysis is handed a universe, a state, a transition probability function, a unit grain, and an update interval, and marginalizes everything outside the candidate system as background (Supplement S3.17). An intrinsic perspective enforced by extrinsic stipulations about what to hold fixed is not intrinsic, and maximal existence is a tie-break among candidates someone selected. A rule for ranking partitions is not a bearer. Universality is not more mathematics; it is a determinant the analyst cannot choose (Doerig et al., 2021).

The cardinality problem then takes a form that can be counted rather than argued. Because the search is recursive, the number of subjects is an output of the maximization: one of the

framework's six-unit examples returns four single-unit complexes, which its authors call monads, alongside one two-unit complex, and in the brain non-overlapping complexes are held to coexist, splitting after callosotomy and possibly during dual-tasking (Albantakis et al., 2023; Tononi et al., 2016). Each is a count of subjects per skull returned over a description that was chosen. IFT counts differently and therefore predicts differently: one body with one set of viability conditions is one policy-integrated persistence regime, so callosotomy divides content and access without dividing the bearer, which is what complete callosotomy appears to show (Pinto et al., 2017; for the dispute, de Haan et al., 2020). The dispute is live, it has a second natural experiment in craniopagus twins who share content across a thalamic bridge while remaining two agents (§12), and IFT is on the falsifiable side of both (Supplement S3.17).

4.3 The Partition Problem and the Interpretation Problem

Every other closure chooses a number or a taxonomy, and the choice has the same shape wherever it is made. Markov blankets are candidate partitions, not physical individuals (Bruineberg et al., 2022), and in canonical nonequilibrium systems cyclic coupling creates dependencies across the nominal boundary, so the blanket degrades as the system becomes recurrent and dissipative (Aguilera et al., 2023). Information Closure Theory makes consciousness non-trivial informational closure relative to an environment, but its authors concede that environments and coarse-grainings are non-unique and that two closed processes yield a closed composite, so the formalism cannot decide whether there are two consciousnesses, one, or three (Chang et al., 2020). Autopoiesis individuates a bounded unity without reaching the conscious level, since bacteria sense, remember, communicate, and regulate without a mind (Damasio, 2018), and Maturana himself denied that coupled autopoietic units thereby compose a higher-order subject, treating society as a medium whose members conserve separate autopoiesis (Maturana, 1980; Maturana & Varela, 1980; Supplement S3.2).

Call the general failure the partition problem: informational, statistical, and causal criteria must be given the cut they score. Blankets require a conditional-independence partition; closure requires a coarse-graining and an environment; IIT searches for a maximal complex over elements and grains and excludes overlapping alternatives (Albantakis et al., 2023). Optimization can make the search exhaustive without making the system description self-individuating. Maturana (1980) diagnosed the general form: distinctions are operations of an observer, and regulation, control, and function are referential relations stated from a metadomain.

The partition problem has a deeper form, the interpretation problem, and it is now a theorem rather than a complaint. Biehl and Virgo (2023) show that beliefs, hidden states, rewards, and even actions can be assigned to a physical machine by an interpretation, including an output-free machine whose "actions" affect only the interpreted world. Fuller et al. (2026) prove that a resource can be bundled with the agent or with every environment without changing resource-relativized Legg–Hutter performance. Neither competence nor optimality decides

whether memory, tools, or computation belong to the agent or merely surround it. Not only the cut and the objective but the resource boundary can be assigned from outside without changing anything the system does (Supplement S3.19). Cardinality, minimal implementation, individuality, and interpretation are one wall.

4.4 The Derivation

So the cardinality question must be answered, and the answer has to be derived rather than chosen, or it inherits the defect it is meant to cure. Here is the derivation, in five steps, and it is the core of this paper.

First, experience is intrinsic. There is something it is like to be the system for the system, not for an observer (Nagel, 1974); IIT is right to make this a postulate (Albantakis et al., 2023), and nothing in the field disputes it. A theory of consciousness is therefore a theory of a bearer, and whatever fixes which physical process is the bearer must be a fact about that process.

Second, it follows that the individuation of the bearer cannot be observer-relative. This is Kanai and Fujisawa's (2024) universality requirement, read as a constraint rather than a desideratum: a criterion that returns a different bearer when the analyst redraws the partition, regrains the units, reselects the consumer set, or reassigns the objective has not identified a bearer. It has identified a description.

Third, every informational, statistical, causal, or functional criterion in the field is observer-relative in exactly this way. That is what the partition and interpretation problems establish (§4.3). The failure is not a defect in any formalism; the formalisms are careful. It is that they are formalisms of ascription, and ascription is what a constitutive theory of subjecthood cannot rest on.

Fourth, exactly one kind of boundary is not drawn by an observer: the boundary a process produces and must keep producing in order to continue existing. This is Maturana and Varela's (1980) autopoiesis, and its formal descendant is organizational closure, a regime of constraints each of which is maintained by the others such that the whole maintains itself (Montévil & Mossio, 2015; Moreno & Mossio, 2015). What makes such a boundary observer-independent is not that it is easy to find but that the system's own continuation depends on it. An analyst may draw a Markov blanket anywhere, and nothing happens to the system when the description changes. An analyst who dissolves a self-produced boundary does not redescribe the system but ends it. Dennett's (2003) Life-world glider is the substrate-neutral illustration: one pattern persists through complete turnover of its pixels, and at the design level it can be damaged, repaired, and killed. The boundary is a fact about the pattern because the pattern's existence depends on it.

Fifth, self-production fixes the kind of individual but not the level, and the level is where the cardinality question bites. Cells self-produce inside organisms, organisms inside colonies, and a theory that counted every autopoietic unit as a subject would be panpsychism of a different

church. What fixes the level is policy closure. A multicellular animal has one body, and one body cannot enact incompatible actions at once; its viability conditions are therefore the conditions of one thing, and the demands they generate (hunger, thirst, suffocation, injury, threat) must be resolved into one commitment because there is one next act. Competing demands that terminate in one enacted commitment are the signature of a single failure mode, and a single failure mode is what "one" means here. A cell inside the organism does not have it, because a cell's demands are not resolved against the demands of other cells into one act; the organism's are. Colonies and corporations do not have it, because their members retain separate fates (Ginsburg & Jablonka, 2019; Maturana, 1980; Supplement S3.7). The level at which boundary maintenance, world-modeling, valuation, and action selection close into one failure mode is the level of the subject.

4.5 The Thesis, and What It Does Not Claim

> Exactly one kind of closure is non-arbitrary: the consumer set closes at the level of a single policy-integrated persistence regime, namely the processes that jointly maintain one active token and resolve its competing demands into one policy. This fixes the bearer of stakes, not the range of beneficiaries that bearer may represent.

Two things should be said about what the argument delivers and what it does not. It delivers the bearer and, with it, the second argument of "recruitable by": the consumers whose standing access makes a content conscious are the processes that constitute this regime's policy closure, and no others, because they are the only consumers whose membership is fixed by the system rather than by the theorist. A content is in the format when those processes can recruit it as structure (§3); the number of contributing modules is then an empirical fact about a particular animal or machine rather than a parameter of the theory. It does not follow that every part of the field is presently recruited, or that every maintenance process is conscious: local circuits may regulate and calibrate the organism while their representations remain C0. Availability-to-many-modules is a symptom, not the criterion.

The argument does not deliver experience from the bearer alone, and it is not meant to. A bacterium satisfies the Indexing Thesis and is not conscious, because it has no format with relational structure for its regime to recruit. An anesthetized body satisfies it and is not conscious, because the format's capacity has collapsed while the organism persists as one regulated token with the same viability conditions, assisted or not. The index answers the question for whom; the format answers what it is like; consciousness is the conjunction, which is why the identity in §1 has two halves and why Mudrik et al.'s (2025) two conditions are the right number. Neither half is sufficient, and the cases that show it (bacterium and anesthesia on one side, cerebellum and neuromorphic chip on the other; §5, 9) are the cases a one-factor theory cannot hold together.

"One policy" means neither one controller nor conscious deliberation. Dennett (1991) is decisive: bodies and elevators can implement policies and meta-policies without awareness.

IFT's criterion is not policy-following but closure: incompatible demands generated by one persistence regime terminate in one enacted commitment because one active token has one next act. The work may be distributed across the loop and across time; integration names closure, not a command headquarters (Dennett, 2003; Supplement S3.7). Closure also requires separating bearer from beneficiary. A policy may represent family, institution, colony, or cause as beneficiary; its bearer is the individual whose maintenance and action selection form one failure mode, which is Ginsburg and Jablonka's (2019) level clause for multicellular individuality: shared fate, aligned component interests, organism-level control, and a boundary maintained by the whole (Supplement S3.7).

Two programs press the level question from opposite ends, and IFT answers both. Levin's group asks what lets cells integrate their activities into a higher-level agent whose goals belong to it and not to its parts, and answers by scaling goals through homeostasis (Pio-Lopez et al., 2023; Safron et al., 2023). That is how a new policy-integrated persistence regime is assembled; subjecthood arrives when competing demands close into one failure mode, not when signaling becomes multi-scale. From the other end, developmental active inference locates the human index before it is autonomous, with homeostasis as the first prior established under co-embodiment (Santaguida & Bergamasco, 2023). Scaffolded is not stipulated: dependence on an external regulator does not defeat an index, and neither does vulnerability to another's decision, since a ventilated patient and a prey animal are both at others' mercy and both have everything to lose. What makes a service a service is that it has nothing of its own to lose: stopped and restored elsewhere, nothing has ended (Supplement S3.20).

4.6 Neighbors

IFT has neighbors at this point, and it should say what it takes from them and where it parts (Supplement S3.24), because the derivation above will be recognized by several traditions as half theirs. The enactive tradition reached self-produced individuality first and added what autopoiesis lacks: adaptivity, the capacity to regulate one's own viability with respect to its limits, from which norms, sense-making, and agency are held to follow (Di Paolo, 2005; Thompson, 2007; Weber & Varela, 2002). Barandiaran et al. (2009) define an agent by individuality, interactional asymmetry, and normativity; Di Paolo's (2009) precariousness, the fact that a living system's processes would run down without its own activity, is what §9 calls constitutive mortality; and Froese and Ziemke (2009) drew the consequence for artificial systems. IFT accepts all of this as the account of the bearer and refuses the step the tradition takes next, which is to treat sense-making as proto-mentality continuous with consciousness. A bacterium is a paradigm sense-maker and has no world. The gradient from bacterium to primate runs through the format, not the index, and the enactive literature has no account of the format and has never engaged the consumer-set problem that availability theories face. Organizational closure (Montévil & Mossio, 2015) is the formal version of the fourth premise and takes no position on consciousness.

Seth's beast-machine account grounds selfhood and experience in the predictive control of physiological condition (Seth, 2021; Seth & Tsakiris, 2018), and it is the nearest neuroscientific neighbor. IFT agrees that regulation is where the subject comes from and disagrees that interoceptive inference is what experience is made of: the content of experience is mostly not about the body, and inference is a service a replaceable system can run. Kriegel's (2009) self-representationalism locates the for-me-ness of experience in a state's representing itself; IFT agrees that for-me-ness is constitutive and denies that it is a representation (§7). Each neighbor has one half. The enactivists have the bearer without the format; the geometry program of §6 has the format without the bearer; the beast-machine account has the right anatomy of stakes and the wrong identification; self-representationalism has the right datum and the wrong relation. The closest engineering precursor is Man and Damasio's (2019) risk-to-self proposal, which ties behavior to the machine's own continued existence so that the machine rather than its owner becomes the locus of concern; IFT accepts the debt and adds the individuation rule it lacks (Supplement S3.8).

4.7 Three Consequences

Three consequences follow immediately, and each is a substantive commitment rather than a restatement.

First, valence is the signature of indexing. Damasio's proto-self is the organism's continuously updated map, and the core self a second-order, wordless account of that organism being changed by an object (Damasio, 1999). IFT's dividing line is exact: the core self is a protagonist in the model; the constitutive owner is the regime whose deviation becomes a demand. The represented self is content; the body–brain loop gives it a bearer (Damasio, 2010). Solms (2019, 2021) supplies the control step: ordinary homeostasis is C0, whereas affect selects deviations into flexible action under uncertainty. Because one body cannot enact all corrections at once, competing needs must be contextually ranked rather than collapsed into one reward currency; felt quality is that categorization, and a scalar reward is not a poor index but the wrong kind of object (Supplement S3.1, S3.14). Damasio (2018) sharpens the continuity physiologically: viability afferents are exposed to their chemical surroundings and relay through barrier-free ganglia, so the process that measures the body is partly continuous with what it measures. A designated safe region is a report produced elsewhere; an indexed viability state is carried by the substrate whose deterioration is at issue, which is what System B of Protocol F4 must achieve.

The prediction that falls out is already in the data. In insular cortex the response to a food cue flips with physiological state: cue-evoked activity present in hungry mice is abolished by satiety, and the cortex represents the future bodily state the cue portends rather than the cue (Livneh et al., 2017, 2020). The same stimulus, with the same sensory topology, is valenced or neutral according to its relation to the stake. Downstream, parabrachial CGRP neurons function as a general alarm that converts heterogeneous threats, from pain and itch to visceral malaise and hypercapnia, into one aversive teaching signal (Campos et al., 2018;

Palmiter, 2018), which is what the ranking of incommensurable demands at a point of policy closure should look like. F3 asks whether the converse is ever true: whether rearranging the topology can invert the sign with the stake held fixed.

Second, lesion evidence against corticocentrism supports indexing. Merker's (2007) upper-brainstem architecture joins motivational ranking, target selection, and action selection; Solms (2019, 2021) locates the convergence in a mesodiencephalic decision triangle centered on the periaqueductal gray and the superior colliculi. Damasio's (1999) coma–locked-in contrast supplies the anatomical knife edge, and Parvizi and Damasio (2003) mapped it: nearby brainstem lesions can abolish the subject or leave it conscious but almost unable to act. IFT divides the labor: cortex and other modeling systems elaborate content, while this regime is the candidate index resolving competing viability demands into one commitment. Extensive cortical damage can remove contents without abolishing the subject (Merker, 2007; Shewmon et al., 1999); focal upper-brainstem damage can do the reverse. The same architecture is identifiable outside vertebrates: Barron and Klein (2016) argue that the insect central complex integrates sensory and spatial information with internal state to select action and is functionally analogous to the vertebrate midbrain, and IFT accepts the analogy and its conclusion, adding that insect experience will be as rich as the recruitable topology and no richer. This is anatomical specificity, not a brainstem theater, and F2 is designed to destroy it if wrong (Supplement S3.6). LeDoux and Brown contest the localization with a higher-order cortical account, and IFT accepts every item they cite: the amygdala is a lower-order survival circuit, not the index (Brown et al., 2019). The disagreement concerns the higher-order assembly, and it is exact: whether re-representation constitutes experience or elaborates a format whose bearer is fixed elsewhere (Supplement S3.1, S3.13).

Third, and this is the discriminating consequence, the indexing relation cannot be stipulated. Historical design is not disqualifying: organisms are products of selection, artifacts of engineers, and the issue is synchronic constitution. Designers attempt to install what IFT says cannot be installed by exactly two routes, and both fail in the same place. The first route installs motivation: appraisal architectures that stabilize drives near a designated safe region, and formal models that derive apparent valence from signed movement along an interoceptive energy landscape (Lian & Goertzel, 2026; Sennesh & Ramstead, 2026). Both are serious control architectures, and neither supplies an internal milieu merely by representing one: violating a safe region or moving down a modeled gradient lowers a score unless it damages the active system as that system (Supplement S3.3, S3.22). The second route installs a self. Bongard et al. (2006) showed that continuous self-modeling supports resilient recovery after damage; Zeng et al.'s (2025) BriSe AI builds five hierarchical self-models in spiking networks, from bodily self and mirror recognition through theory of mind, affective empathy, autobiography, and narrative identity (Supplement S3.10). The achievement sharpens the failure: every level represents a self, and the index is not a representation. An externally powered, fungible system can model its body, attribute beliefs, mirror distress, and narrate its history while remaining a service. The oldest version of the claim, that a world model

compressing its own history recovers the agent as its most persistent invariance and is thereby self-aware, is reported and not endorsed from inside the world-modeling program itself (Safron, 2023); an invariance recovered from a data stream is the cheapest encoding of a regularity, not a stake. Neither predictive error nor technically "intrinsic" motivation supplies the fate: predictive-coding residuals organize perception without being bad for the system (Rao & Ballard, 1999), and variational empowerment makes a universal agent seek high-control states with no immediate reward advantage (Hayashi & Takahashi, 2026; Supplement S3.4, S3.5). Labeling a negative scalar "pain" does not make it hurt; the issue is what the running system must preserve to remain itself.

This does not require carbon, metabolism, or human physiology, and artificial valence could be alien in content and implementation. A designer may build the system but cannot confer the index by declaration, by assigning reward, or by naming a variable "homeostasis." Non-stipulability means the persistence conditions are enacted by the running system and are token-specific: changing them changes or destroys the individual rather than editing its task.

4.8 The Identity at Work

It is worth seeing the identity classify the canonical cases in one place, because a theory that is one sentence long should be checkable in one paragraph. Bacteria: index without format, unconscious. Deep anesthesia: index without format, unconscious, with PCI recording the collapse of the format while the token and its viability conditions persist (§6). Dreaming: format partly instantiated, regime intact, motor output gated, so dreams are felt, valenced, and organized around the stakes of the dreamer, and dreaming survives the destruction of the REM generator because it was never the state-setting mechanism that owned it (§8). Cerebellum and enteric plexus: format-like structure not recruitable by the regime, unconscious (§5). Blindsight: the topographic field is destroyed in the scotoma, so what survives through extrastriate routes is diluted content that biases choice without being recruitable as structure (§6). Split brain: one regime, divided content, one subject (§4.2). Craniopagus twins sharing a thalamic bridge: two regimes, shared content, two subjects (§12). Insects: a small format recruitable by a central complex, conscious at the scale of the format (§4.7). A neuromorphic chip with cortex-like topography, externally powered, compiler-scheduled, and fungible: format without index, unconscious (§9). A language model: neither a standing recruitable topology nor a regime, and the wrong kind of thing on both counts (§10). The classifications are not a list of intuitions the theory was fitted to; several of them are the predictions §12 puts at risk.

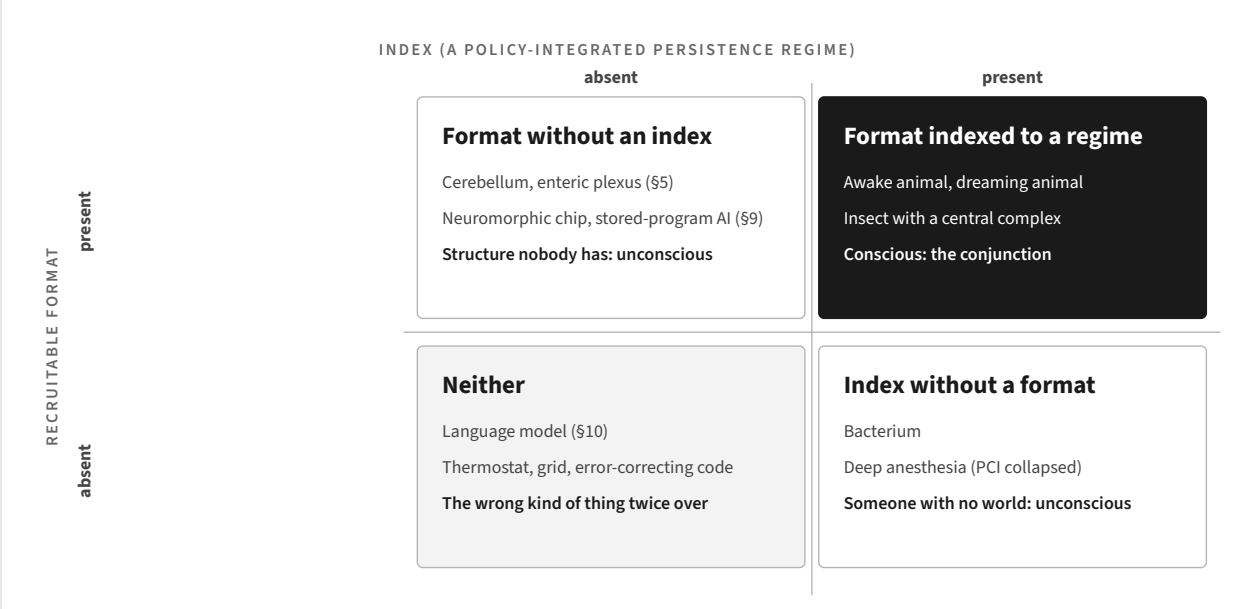

Figure 2. The identity at work (§4.8). A recruitable format and a policy-integrated persistence regime are each necessary and only jointly sufficient. Every cell but one is unconscious, and the cells are the predictions §12 puts at risk.

5 The Cerebellar Test, and What It Eliminates

The single most underused constraint in consciousness science is the cerebellum, and any theory that cannot pass it should be discarded.

The cerebellum contains more neurons than the rest of the brain combined and is not the feedforward controller it was once caricatured as being (De Zeeuw et al., 2021). Its microcircuits are recurrent at every level, its plasticity is distributed across synapses and timescales, its climbing fibers carry reward-, prediction-, and success-related signals, and its memory traces migrate during consolidation (Supplement S1). Yet cerebellar destruction, devastating for coordination and many cognitive functions, does not abolish the conscious subject. The cleanest case is a living one: a woman with complete primary cerebellar agenesis, discovered at age 24, with moderate ataxia and dysarthria and an otherwise ordinary life, including marriage and a child (Lemon & Edgley, 2010; Yu et al., 2015). Nobody has proposed that she is not a subject.

The inference is stark. Recurrence, plasticity, reinforcement-like learning, error signaling, temporal credit assignment, enormous neuron count, and dynamically stored memory are jointly insufficient. This meets Recurrent Processing Theory at its strongest point. Lamme (2006) is right that feedforward activation can support sophisticated discrimination without reportable experience, and that frontoparietal recruitment should not be smuggled into phenomenality. But the move from “recurrence marks a visual format” to “recurrence constitutes a subject” fails the cerebellar test, and recurrence is therefore a format-building mechanism rather than a constitutive criterion (Supplement S1).

Every major theory can redescribe the exclusion; almost none derives it without importing its preferred anatomy (Seth & Bayne, 2022). IIT now has the best version and it should be stated. In small networks, weak coupling between modules creates fault lines along which the minimum partition falls, so a modular system condenses into several small complexes rather than one large one; cerebellar micro-zones are offered as the biological instance, and divergence from a dense cortical lattice is presented as a consequence of connectivity rather than a stipulation about cortex (Albantakis et al., 2023; Tononi et al., 2016).

That is the right kind of mechanism, and it still does not deliver the exclusion, for two reasons the framework concedes. No exhaustive calculation exists for a cerebellum, and full analysis is feasible only for idealized systems of a few units (Albantakis et al., 2023; Findlay et al., 2025; Marshall et al., 2026); an analogy from six units is not a derivation for sixty-nine billion neurons (De Zeeuw et al., 2021). More seriously, fault lines do not eliminate; they multiply. Carving a modular system along its fault lines licenses the small complexes it yields, which is why the same analysis returns monads, so the modular reading predicts for the cerebellum not the absence of a subject but a large population of very small ones. RPT owes a non-anatomical reason why recurrent visual content qualifies and recurrent cerebellar content does not; workspace theory restates absent cortical broadcast; informational closure describes a cerebellar forward model almost by definition (Chang et al., 2020). Declining to calculate is not prediction, and multiplying subjects is not excluding one.

Cerullo (2015) showed how cheap redescription is: a deliberately trivial theory reproduced IIT's two advertised verdicts, two complexes after split-brain surgery and no conscious cerebellum (Supplement S1.3). A parody matched the flagship's explanatory record, so the exclusion supports only a criterion fixed independently of the case. IFT's was fixed in §4 before the cerebellum was raised, and extends unmodified to the enteric plexus.

Higher-order theory is the exception, and the credit should be given plainly: cerebellar states are never re-represented by the prefrontal networks it identifies as the site of inner awareness, so their exclusion follows from the theory instead of being appended to it (Brown et al., 2019). That is a real derivation, and it is why §7 treats the higher-order family as this account's serious rival. What it does not deliver is the bearer: a higher-order representation is a representation, and the question of whose awareness it constitutes reopens the cardinality problem one level up.

IFT answers with a general elimination rule that is the membership criterion of §3 applied. The cerebellum's recurrence is within-model recurrence: it calibrates forward and inverse models, and its internal states are consumed by effector loops and cortical circuits as corrections. Those outputs serve the indexed organism without the cerebellar representations being recruitable, as structure, by the processes that close the organism's policy; being useful to a conscious system is not the same as being a conscious representation in it. The cerebellum is, in Dehaene's (2014) taxonomy, a magnificent engine of latent and diluted information, yet never explicit in the policy-integrating format.

The cleanest behavioral test is one learning problem in two variants. Delay eyeblink conditioning survives decerebration, amnesia, and distraction and needs no awareness of the contingency; insert a trace interval and the task recruits forebrain structures and, in humans, tracks awareness of the predictive relation (Dehaene, 2014; Supplement S1). Temporal bridging is not the criterion; it marks where format-eligible content becomes necessary.

The enteric nervous system is the harder control because it is not subordinate to cortex: several hundred million neurons, overwhelmingly intrinsic, recurrent, largely autonomous, and a major tributary to feeling (Damasio, 2018; Furness, 2012; Supplement S1.2). Its afferents enter the indexed format, which is why gastrointestinal states are felt; its own computation coordinates peristalsis without becoming an experience. Being a tributary to a format is not membership in it, and the plexus was never the bearer. On the modular reading it is not excluded but promoted: an autonomous, recurrent, non-subordinate network is a candidate second complex, and a theory that condenses a body into complexes owes an answer about the one in the abdomen.

Hence the rule: recurrence whose contents are never recruitable by the indexed policy format is metabolically useful and phenomenally inert. This is not an ad hoc patch, because it also disposes of the strongest formal counterexamples to integration-based accounts. Aaronson (2014) constructs linear transformations over finite fields that maximize effective information across bipartitions, and Merker et al. (2022) argue that Φ-like measures may track efficient global information transfer, found in cortex but also in power grids. Both objections take the resolution the cerebellum and the enteric plexus force on biological grounds: an error-correcting code has maximal integration and no self-constituted predicament.

6 The Geometry Thesis: Quality Is Format Topology

The structure question, why an experience has the character it does, has been widely regarded as untouched. That is no longer true, and the result that changed it should be extracted from the theory that produced it.

Damasio (2018) states the representational constraint: a neural map need not resemble a picture, but must preserve relations among its parts. Haun and Tononi (2019) formalize this for spatial extendedness as four relations among overlapping “spots” (reflexivity, connection, fusion, inclusion), from which region, location, size, boundary, and distance follow. Unfolding an eight-unit line grid, they find 133 irreducible distinctions among 255 candidate mechanisms, 117 of the 133 fully extended. The decisive control is random rewiring: with the same number of connections it yields comparably many distinctions and virtually no extendedness. Phenomenal structure tracks topology, not activity or information content. Lesions to topographic grids accordingly produce scotomas in which a region of space ceases to exist rather than appearing dark, and altered lateral connection strength warps experienced distance without changing firing rates.

Two results from the same program sharpen what a standing format is and what F1 must perturb. A merely inactive unit still contributes: switching it off changes which structure is specified rather than removing structure. An inactivated unit (fixed, with no counterfactual states) collapses every distinction and relation it entered, and the substrate itself shrinks (Albantakis et al., 2023). The prediction is a decade old: changing connection efficacy within the substrate should change experience even when activity does not, and a silent area can contribute to an experience while an area whose connections are cut cannot (Tononi et al., 2016). Standing organization, not activation, is the carrier.

The 2025 extension makes the consequence explicit. Object and scene content presupposes mid-level feature groupings, and those presuppose a spatial field of nested, overlapping regions and relations, most of which can remain unrecognized, unremembered, and unreported while still constituting what seeing is like (Haun & Tononi, 2025). The Geometry Thesis therefore entails phenomenal overflow (Block, 2011): a sparse functional summary can ride on a vastly richer format, but it cannot substitute for it. A structure also cannot be transmitted through a limited-capacity report channel, so phenomenal capacity exceeds access capacity (Tononi et al., 2016), and partial-report paradigms already show a large-capacity store that outlasts what is reportable (Landman et al., 2003). Coordinates, object labels, or a compressed scene description may support competent behavior while failing to instantiate the relational field they summarize.

The overflow debate has stalled because the dispute is over what is not reported, and report cannot settle it (Cohen et al., 2016). There are two serious alternatives to standing first-order geometry. Higher-order theorists place the impression of peripheral detail in misrepresentation at the higher-order level: subjective inflation (Knotts et al., 2019; Odegaard et al., 2018). Inflation is not ad hoc: it predicts elevated false-alarm rates for peripheral and unattended detection and survives the response-bias objection (Brown et al., 2019; Supplement S3.13). Perera (2026) offers a different denial: a sparse query service can virtualize a dense personal-level field, making arbitrary content operationally available without maintaining a saturated relational representation, so richness is real as a functional abstraction and absent as standing substrate. Mudrik et al. (2025) deny a third thing: unaccessed content is not phenomenal at all, because access is a condition on consciousness rather than a consumer of it. The four positions separate cleanly. Inflation locates richness in higher-order misrepresentation; virtualization in content-on-demand; the two-conditions account says unaccessed structure is not yet conscious; IFT says unqueried relations are standing constituents of a conscious field (Supplement S3.21, S3.23).

They are not empirically equivalent, and this is the point at which the stalemate ends. Perturb never-recruited posterior relations at their source rather than asking about them. IFT predicts a later relation-specific error even when an independent post-cued assay shows the query-generating service intact; inflation predicts no effect; virtualization predicts an effect only to

the extent that the service is impaired; and the two-conditions account predicts one only where access follows. Protocol F1 is that four-way experiment.

I take the spatial derivation and its 2025 extension to be the strongest positive program in consciousness science. Tononi (2004, 2012) had already stated the geometric ambition, but Haun and Tononi (2019, 2025) supply the relational derivation that can be separated from IIT's metaphysics. IFT keeps standing topology capable of realizing phenomenal geometry and rejects the identity of consciousness with Φ, the partition-relative maximization, exclusion over nested subjects, intractable exact computation, and vulnerability to unfolding (Cerullo, 2015; IIT-Concerned et al., 2025). Safron (2020) proposed the same surgery from inside the free-energy camp; IFT's grounding is the index, not cybernetic control.

The surgery also reclassifies the field's most successful clinical measure. Perturbational complexity, though introduced inside IIT, is compatible with global workspace theory (Casali et al., 2013; Farisco & Changeux, 2023); on IFT it measures the capacity of the standing format, upstream of whichever consumers recruit from it, which is why it indexes state and is silent about content (Supplement S4.2).

Hanson and Walker (2021) make the unfolding challenge exact: a theory is pre-falsified if a transformation leaves the inferred conscious state fixed while changing its prediction. IFT escapes only if indexing belongs to an independently fixed inference basis, so removing token-specific persistence coupling must change the evidence and not merely the verdict. A surrogate reproducing outputs without the original's conditions of failure does not hold the IFT inference fixed; one given the same constitutive vulnerability is a second indexed system. Protocol F4 preregisters index evidence before comparing dissociation manifolds (Supplement S2.6).

The topology at issue is indexed topology, not bare graph structure. Damasio's (2018) division between the musculoskeletal frame and sensory portals, which establish perspective, and the old visceral interior, which supplies the valenced condition of life, is the biological antecedent of T3 and T2. Dennett's (1991) color analysis adds the ecological constraint: quality spaces are defined relative to discriminators that coevolved with ecological differences, so no observer-independent red is there for a network simply to mirror. T3 identifies quality with the relational organization of distinctions for an indexed consumer regime and separates quality from valence. Topology can therefore specify color, pitch and timbre, arousal, bodily location, and modality while leaving sign underdetermined; sign is the format's relation to the system's persistence (Supplement S3.12, S3.14).

The two theses together predict the established separation of pain structure from pain affect, in both directions. Frontal intervention and hypnotic suggestion can leave sensory location, quality, and intensity intact while removing unpleasantness and changing cingulate rather than primary somatosensory activity (Damasio, 1999; Rainville et al., 1997); pain asymbolia is the lesion form (Berthier et al., 1988); and a postcentral lesion can leave pain affect without pain sensation (Ploner et al., 1999). Structure and sign are separately manipulable, and

bilateral destruction of the insula leaves feeling and sentience in place (Damasio et al., 2013), which is what a theory should expect if the stake, not the map, carries the sign. Protocol F3 tests the missing converse: whether reorganizing interoceptive topology while holding the physiological stake fixed can invert hedonic sign with full reinforcement, exploration, and memory consequences. If it can, the Indexing Thesis is redundant.

7 Metacognitive Typing and the Origin Thesis

The fourth factor is the system's model of its own signals, and it is the layer at which the hard problem is generated.

Lau's (2008) higher-order Bayesian formulation states what this layer does. Conscious confidence depends not only on first-order signal quality but on the system's fallible model of how its own signals behave, so blindsight, hallucination, and dreaming may result partly from misestimated higher-order models rather than absent first-order information. Graziano's (2022) Attention Schema Theory supplies the mechanism for awareness reports: a schematic model of selective attention, omitting the machinery that implements it, portrays attention as a nonphysical possession of an object (Supplement S3.13).

Broadcasting and higher-order modeling are not competitors: access explains how a content acquires broad causal influence, typing how a system assigns epistemic status to its own processing. IFT adds only that typing is a service to the policy. That service can be learned rather than built (Cleeremans et al., 2020), and learning it changes nothing constitutive, because learning to care presupposes a bearer for whom the learned value is not merely a score (Supplement S3.11).

Higher-order theory is the one family that makes this layer constitutive, and it is IFT's closest competitor. Its two commitments are that a first-order representation is never sufficient for experience, and that a state one is in no way aware of being in is not a state there is something it is like to be in: the Transitivity Principle (Brown et al., 2019). Three misreadings should be set aside, because IFT has no interest in inheriting them: higher-order theorists do not equate consciousness with metacognition, do not require sophisticated thought, and are not committed to consciousness scaling with prefrontal activity (Supplement S3.13).

The disagreement is over transitivity, and it is exact. IFT denies that awareness of being in a state is a condition on that state's being phenomenal, and replaces it with standing recruitability within the indexed field. Nothing else in this paper survives if transitivity holds: overflow is false if nothing unnoticed is phenomenal, the Geometry Thesis loses the structure it derives, and F1 has nothing to find. Conversely, transitivity buys two things this account has to work for: a derivation of the cerebellar exclusion (§5) and an explanation of peripheral richness (§6). The positions make opposite predictions under perturbation, and one of them is wrong.

Why IFT breaks with it is why it breaks with availability accounts. A higher-order representation is a representation, and the question it was introduced to answer returns: awareness to whom? If a first-order state becomes conscious by being meta-represented, and the meta-representation is not itself something one is aware of, what makes the pair one subject's is unspecified, and a further level defers the question. This is the cardinality problem relocated: the consumer set is closed at one distinguished consumer instead of left open, a gain in determinacy and none in principle. Kriegel's (2009) self-representationalism is the version closest to IFT, because it replaces monitoring by a distinct state with a state's representing itself, and the disagreement is the same one: unowned experience is indeed incoherent, but what the transitivity intuition detects is the index, not a representational relation, which is why a state can belong to someone without anyone or anything representing it (Supplement S3.13).

The physiological form of annotation is the mechanism that carries the format. If format membership is standing modulatory availability (§3), typing is the setting of gain: confidence is the weight with which content is permitted to act. Solms (2021) identifies this precision adjustment with consciousness. IFT keeps the mechanism and refuses the identity, because a system for which nothing is costly can annotate reliability indefinitely without the annotation meaning anything. The assignment must be made by whatever bears the cost of assigning it wrongly (Supplement S3.6).

Now the residuum. Frankish's (2016) illusionism denies irreducible phenomenal properties and replaces the hard problem with explaining why introspective systems judge that they possess qualia. Dennett's (2018) no-double-transduction argument supplies the mechanism: consumers receive representational content, never the medium, and project features of content onto an imagined substrate. Metzinger (2003) named the general property phenomenal transparency: a representation whose vehicle properties are unavailable to introspection is lived as the world rather than as a model, and the transparency of the self-model yields the impression of an irreducible self.

I accept the mechanism and reject the verdict. Here is the Origin Thesis.

A format is transparent to its users by construction. Consumers receive content; there is no channel by which they could receive the medium, and no transducer that would convert medium into content. A system modeling its own modeling therefore faces a structural asymmetry: it can represent what its format represents, but the format's own constitution is not among the things the format can hold. When it asks what its states are made of, the only representational resource available is content, and content will not answer. The result is a stable, ineliminable impression of something present and unspecifiable.

Illusionism stops here and calls the impression a mistake, and Metzinger stops here and calls the self a model with no one behind it. The structural point predates both formalizations. Merker (2007) described the conscious field as a body-within-world coordinate system whose "ego-center" is its origin: objects are apprehended from it, but the origin cannot become an

object within the field. The projective consciousness model generalizes the insight in three-dimensional projective geometry, representing the singular viewpoint as an origin outside represented space (Merker et al., 2022; Rudrauf et al., 2017; Williford et al., 2018). Neural implementation remains unresolved and I claim nothing about it. What matters is the theorem: the origin of a projection is not an element of the projection. Where IFT parts from Metzinger is on what transparency hides. On the self-model theory there is no one, because the self is a model; on IFT there is a bearer, because the regime is not a model, and what transparency conceals is not a fictitious self but the constitution of the format whose contents the bearer recruits.

Two independent formalisms converge. Lahav and Neemeh (2022) argue that phenomenal states are direct only within the cognitive frame whose dynamics they drive; elsewhere they are physical patterns. Perera (2026) reaches the same result in implementation language: sparse subpersonal machinery delivers a dense personal-level world across an abstraction boundary that hides production from its consumer, yielding de facto rather than metaphysical privacy. IFT accepts the structure and rejects the conclusions. Frame-relativity generates rather than dispels the residue, and rich self-simulation does not fix a bearer because it remains representational content plus metacognitive typing (Supplement S3.15, S3.23).

The intuition of an inexplicable further fact is therefore veridical about structure and misleading about ontology. The origin cannot appear within the field; the mistake is to infer an extra property. Illusionism is right that nothing must be added, anti-illusionism that the intuition is not confusion, and Perera that privacy is architectural. They describe one geometry.

This is not an argument from bafflement. Churchland (1996) is right that failure to imagine a neural mechanism carries no metaphysical weight and that progress requires rich reductive contact across levels. Dennett's (2001) topic-of-consciousness question supplies the complementary constraint: even a new sentience-bearing physics would not explain how myopic parts compose a system about Braque or Kilimanjaro. T4 accepts both. It predicts a representational limit inside a fully mechanistic explanation, not a limit on mechanistic science. The Phenomenal Concept Strategy is its closest precursor: a projective format can present what it is like without presenting how that format is constituted (Balog, 2012; Supplement S3.16).

Damasio (2018) and Solms (2019) supply the biological half: feeling qualifies organismal states by value and selects unmet needs for action. §4 takes the function, anatomy, and stakes. Neither explains why the completed control story remains opaque to its user; T4 supplies that missing half without adding a property.

Explanatory satisfaction means deriving phenomenal structure from substrate, and spatial extendedness shows that this is achievable. Metaphysical satisfaction means making it unintelligible that the organization could obtain without experience, and IFT predicts that

demand cannot be met: the same transparency that makes a format usable makes its origin unavailable as content. Chalmers's (1995) easy problems are answered by the format and its consumers, Aaronson's (2014) Pretty-Hard Problem by the index, and the hard problem by that structural limit (Cerullo, 2015). The hard problem is neither confused nor a prompt for new physics; it is a predicted artifact of the architecture that has it.

The practical payoff is that avowal closes symmetrically. An indexed system with metacognitive access should generate hard-problem intuitions, and a system trained on human discourse can generate the same reports without a format (Butlin, Bayne, et al., 2026). Insistence on an ineffable remainder is therefore predicted of a conscious artifact and trivially imitable by a language model, while silence is equally weak because an indexed system may lack language. A prompt can make one model assert, deny, or hedge (Porębski & Figura, 2025), so avowal policy is no evidence of an index (Colombatto & Fleming, 2024; Supplement S5.4).

The Origin Thesis is the least constrained part of the account. It explains why the hard problem is generated and why metaphysical satisfaction may remain unavailable, but the projective formalism currently functions as a philosophical completion rather than as a comparably mature empirical result. T1–T3 do not depend on accepting it.

8 Why the Cogitate Result Has the Form IFT Expects

An adversarial collaboration in which both parties lose is usually read as a setback. On IFT, the pattern follows from the distinction between the format that carries content and the consumers that recruit it, and it could have been written down before the data came in. Safron's (2022) IWMT anticipated the same posterior/frontal division; IFT adds that neither region nor integration individuates the bearer (Supplement S3.9). Mudrik et al. (2025) draw the same lesson, tabulating which theories supply a content mechanism and which an access mechanism: the Cogitate outcome written as a division of labor. IFT supplies the column their table does not have.

Sustained posterior content. If format geometry is realized in standing topographic connectivity, the substrate specifying conscious content is posterior and should remain decodable as long as the content is present. Cogitate found this across three recording modalities, and the orientation result is the more diagnostic: a feature that was always task-irrelevant decoded reliably posteriorly and sat near chance prefrontally. That is what a format looks like: it holds features whether or not any consumer needs them (Haun & Tononi, 2025).

Transient prefrontal engagement. Dehaene et al. (2017) treat high-level ignition as the neural implementation of C1. IFT preserves the functional claim while denying that C1 is identical to the format it accesses: prefrontal cortex is a consumer, so ignition should be brief, task-modulated, and absent when no consumer is engaged. The missing onset-plus-offset profile is what that distinction requires. GNWT's error was never its architecture but the identification

of experience with availability; this is the move the cardinality problem shows to be ill-formed.

The synchrony null was IIT's own doing. Its flagship structural result establishes that phenomenal structure tracks topology rather than activity (Haun & Tononi, 2019), yet the preregistered prediction looked for content specification in dynamic inter-areal coupling. By IIT's own best result, content specification should be a property of standing connectivity, which the paradigm did not manipulate and could not have detected. IIT lost a prediction it should never have made, and the null is not evidence against the Geometry Thesis.

A third framework anticipated the report-dependent half. Whyte et al. (2026) review an active-inference model in which reporting is a mental action that gates second-level precision; under no-report conditions, the model produces reduced prefrontal-like activity and no P3b-like response while retaining a high second-level posterior (Supplement S4). It did not predict the full Cogitate result and therefore did not "win" the collaboration. The methodological point is stronger than the slogan: an adversarial design can eliminate both contestants while leaving a viable explanation outside the contest (Lakatos, 1970).

Sleep science supplies the precedent. REM was promoted from correlate to identity and studied for decades as the objective replacement for dreaming. The identity was wrong: dreaming survives destruction of the pontine REM generator and is abolished by ventromesial frontal and inferior parietal lesions that spare REM, and the driver proved to be mesocortical–mesolimbic appetitive seeking rather than the cholinergic switch setting global state (Solms, 2000, 2021; Supplement S4.1). A marker had replaced the phenomenon, and the correction relocated the mechanism from state-setting to stakes. GNWT identified experience with availability, and Cogitate charged for it.

9 Substrate: Resolving the Ambiguity on Which the Machine Question Turns

There is a single unexamined ambiguity at the heart of the current literature, and nearly every disagreement about machine consciousness inherits from it.

Friston's (2010) free-energy principle unifies the layers described so far and is the shared core of two mutually exclusive verdicts. Whyte et al. (2026) extract a permissive minimal theory identifying conscious contents with sufficiently precise and slow posteriors over hidden states of world, body, or brain. Seth (2025) extracts from the same formalism a restrictive biological naturalism on which digital AI is off the trajectory. The mathematics does not adjudicate, because it does not say whether the free energy that matters is an information-theoretic quantity a system computes or a thermodynamic quantity a system is.

The permissive pole has a sharper statement outside that formalism. Lahav and Neemeh (2022) argue from an equivalence principle between a conscious human and a cognitively

duplicate artificial system that zombies are not nomologically possible. The proof individuates its agents by stipulating agreement on every measurement, then closes the consumer set at one designated subsystem, reproducing the cardinality problem in its strongest form; and its most interesting result cuts the other way, since two systems with identical dynamics at different positions are not one first-person frame. Token individuation is doing the work, and position supplies it without supplying stakes (Supplement S3.15).

The restrictive pole has three versions, and IFT agrees with each about the verdict and with none about the reason. Seth's (2025) case can be read as an individuation argument: exact digital realization of continuous, stochastic, or chaotic processes is unavailable; neural replacement assumes the disputed grain of equivalence; mortal computation collapses the software–hardware separation; and evolved systems are not causally closed at one scale (Hinton, 2022; Supplement S2.1). Biological computationalism holds that brain computation is inseparable from its substrate, so that software layered on interchangeable hardware is the wrong kind of thing however sophisticated the algorithm (Aru et al., 2023; Milinkovic & Aru, 2026). And Findlay et al. (2025) supply the strongest formal result: a stored-program computer exactly simulates a four-unit Boolean system yet has zero system integrated information and fragments into trivial micro-complexes, while the target specifies a rich cause–effect structure, because, in the theory's own words, being is not doing (Albantakis et al., 2023; Supplement S2.2). Their verdict is the negative half of IFT in another vocabulary: a machine that does everything and is nothing. The opposing pole is organizational invariance and substrate-neutral self-maintenance (Chalmers, 1995; Dennett, 1998; Safron, 2020; Solms, 2021); biological realism describes how an index may be realized and does not identify the subject.

IFT converges on that verdict for stored-program machines from a premise set it shares with IIT at no point: two theories disagreeing about what consciousness is nonetheless agree that functional equivalence does not deliver phenomenal equivalence, making the conclusion robust to the failure of either. But the derivations are not interchangeable, and where they part is the discriminating case for both. Findlay et al. rest the exclusion on causal topology, which the unfolding argument attacks and which a kernel author can change by compiling the same model as a recurrence or as a matrix product (Supplement S2.6). The biological computationalists rest it on the grain of the computation. IFT rests it on individuation: interchangeable hardware has no boundary of its own to lose. The escape hatches the formal result leaves open, neuromorphic hardware without a central processor and cerebral organoids with the right anatomy, are therefore cases IFT treats differently from either rival. Klatzmann and Doerig (2026) argue that biological naturalism divides into an untestable claim that biology is intrinsically special and a testable claim that biology enables distinctive information processing, the second compatible with computational functionalism. IFT is neither. Its restriction is not on material and not on processing but on individuation, and it is testable because individuation can be built, which is what F4 does.

One hatch has been fabricated and it closes in the wrong direction. NorthPole eliminates off-chip memory, intertwines memory with compute, and preserves cortex-like topography, yet is deterministic, compiler-scheduled, unable to learn in place, stateless across frames, and exposed to its host through three commands (Modha et al., 2023; Supplement S2.4). High integration without an index is a cerebellum with better press. Organoids are the harder case and not an automatic positive. Brainoware has shown plasticity-dependent reservoir computation in a cortical organoid, and the organoid-intelligence program proposes closed-loop hybrids linking organoids to sensors, outputs, perfusion, and biofeedback (Cai et al., 2023; Smirnova et al., 2023; Supplement S2.3). But living cells enact boundaries at the cellular level, which does not individuate the organoid–computer assembly as one subject.

IFT resolves the ambiguity and gives each side half of what it wants. It breaks with the computational functionalism underwriting the original indicator program (Butlin et al., 2023): the index is multiply realizable but not computationally invariant. Two systems may implement the same algorithm while only one has running states constitutively tied to the continued existence of that active token. Computation can specify the format and its services; it cannot specify which physical process is the subject.

Rejecting computational invariance incurs the fading- and dancing-qualia objections (Chalmers, 1996; Cerullo, 2015), which assume that one subject persists through neuron-by-neuron replacement and force experience to fade invisibly or vanish at an arbitrary count. IFT rejects the shared premise. Maturana and Varela (1980) supply the criterion: structure may change while organization persists. Replacement absorbed by the system's own viability regulation is structural, and nothing fades at any count; replacement that transfers the computing to an externally maintained substrate changes the unity, and a different individual stands on the far side. There is no interval in which a subject exists who notices the change and cannot say so, because in the first regime nothing fades and in the second the entity that would notice has been dissolved rather than silenced (§10; Supplement S2.7).

The constitutive requirement is non-stipulability, not thermodynamics: the system's persistence conditions must be facts about the running individual rather than choices about its task, and the boundary must be enacted by the same active token whose inference and action preserve it (Maturana & Varela, 1980). Current language models are excluded neither because they are algorithmic nor because they run on silicon, but because their operation constitutes no token-level persistence regime. Their energy use makes the point only after reinterpretation: for an organism, energy is acquired, budgeted, and defended by the same policy that models the world, and budget failure is death; for a model it is an operator's externality. The issue is not joules but who is answerable for them (Porębski & Figura, 2025; Supplement S5.1).

Two clarifications keep this from being misread as a claim about technology. The symptom by which a missing index is recognized in artifacts is externally fungible persistence: the instance can be stopped, and service restored on other hardware from a saved state, with

nothing ended, because identity is defined at the orchestrator. The constitutive fact the symptom reveals is the absence of a self-produced boundary at the level of the running process; fungibility is the diagnostic, not the definition. And pausing is not the issue. An anesthetized patient, a hibernating bat, and a desiccated tardigrade are halted tokens, not restored copies, and a future ability to scan and duplicate a brain would create second subjects rather than abolish the first, exactly as a surrogate given the original's conditions of failure is a second indexed system (§6).

Hinton's (2022) mortal computation points in the right direction but supplies the wrong mortality. Device-specific learning makes parameters die with hardware, yet the function transfers by distillation, so what dies is an encoding rather than a subject (Supplement S2.5). Constitutive mortality makes computing part of what keeps this device in existence: the computing substrate is the substrate maintained, energy acquisition and repair are endogenous, and no checkpoint restores the active token. Mortality of substrate and integration of policy are separate achievements, and IFT requires both. Biology has a monopoly of current examples, not of principle.

The anti-exceptionalist objection is that IFT makes consciousness cheap in nature and expensive in industry. But self-maintenance alone is not the criterion. IFT requires a structured format, a policy-integrated persistence regime, and a non-stipulated relation between them; these are conditions common in animals and systematically engineered out of services. No Wonder Tissue is invoked: the restriction is not on material but on what makes one thing whose failure is its own.

10 What It Would Mean for an AI to Be Conscious

The missing requirement is one relation, not another module. The field keeps finding lists because it inherited a question designed to screen the relation off. Turing (1950) did not offer the imitation game as a theory of consciousness; he replaced "Can machines think?" with performance, and modern AGI taxonomies inherit the substitution as a capability program (Morris et al., 2024). The game excluded body, voice, appearance, and touch to keep the test on the intellect, and every one of those exclusions removes a way of failing: a body can be damaged, a voice is produced by machinery that can be injured, presence in a room is exposure. What remained was a teleprinter channel, precisely the channel over which a system with no stakes can be trained to say everything a system with stakes would say. On IFT the inversion is exact: what the game excluded is the criterion, and what it retained is the channel most directly optimized.

This is not an intuition pump against silicon: Dennett (1991) was right that inability to imagine how a robot could be conscious is no argument, and IFT's exclusion concerns the identity of the system, not the poverty or mechanical character of its computation. It should also be separated from two families with which it is regularly confused. Incapacity arguments

hold that organisms find unlistable affordances no algorithm can enumerate, so that general intelligence in machines is impossible (Kauffman & Roli, 2023); quantum arguments locate the unity of experience in entanglement and represent anesthesia as decoherence (Kharchuk, 2025). The first mislocates something real: affordances are unlistable because they are indexed, what counts as a use being fixed relative to a bounded individual that must keep existing, which is individuation, not computational power. The second describes a vehicle: coherence would say how a format is held together, not whose it is, and its showpiece is already covered by the global-state factor without it. IFT declines both halves of both families: nothing here forbids artificial general intelligence, which may arrive unaccompanied, and nothing here requires new physics, since a collapse is an event and an event is not an owner (Supplement S2.8). Instantaneous unity is a fact about a format. A subject is a fact about a fate.

Chalmers (2023) lists the strongest obstacles for text-only models: embodied perception, world and self models, recurrence and persistent memory, a global workspace, and unified agency. These construct the format and its consumers but not the index. Storage continuity, computational continuity, and subjective continuity are different: information can persist, state can depend on prior state, and yet no active token owns unfinished vulnerabilities and expectations. Damasio's profoundly amnesic David shows the converse: autobiographical storage can collapse while core consciousness remains (Damasio, 1999). A resettable avatar therefore has no stakes merely because it has a body schema and memory.

Multidimensional indicator programs make the missing relation explicit by omission. Evers et al. (2025) profile sensory, evaluative, agential, mnemonic, and metacognitive capacities while conceding that externally defined reward limits autonomy; Butlin et al. (2023) operationalize recurrence, workspace, higher-order monitoring, attention schema, predictive coding, agency, and embodiment while denying that the list settles consciousness. It remains an inventory of services without an account of whose services they are (Butlin, Long, et al., 2026). IFT supplies the generator and the missing bearer (Supplement Table S1; S5.1). Cea's (2026) review makes the dispute explicit: dissociation, functional association, or phenomenological association between general intelligence and consciousness. IFT accepts capability dissociation and relocates the association in indexing. Morris et al. (2024) define AGI by performance and exclude consciousness; Legg and Hutter's (2007) universal intelligence supplies goals through environmental reward without asking whether reward is good or bad for the active token. Memory, multimodality, tool use, longer horizons, and agentic training can therefore improve C0, C1, and C2 indefinitely without making the token non-fungible.

Three consequences are worth stating plainly because they invert common expectations.

First, the first conscious artifact will probably be unimpressive. The strongest near-term candidates are small, embodied, physically vulnerable systems whose sensorimotor loops close around real damage. Ginsburg and Jablonka (2019) predict that a conscious robot would be developmentally unique rather than factory-interchangeable. IFT accepts the token-

specificity and rejects development as constitutive: a directly assembled system may qualify if its present operation makes its continuation its own problem. Ma's (2025) boundary-inference program has the right evidential shape, but body ownership remains a self-model until viability reorganizes attention, learning, memory, and action.

Second, we will not believe it. A small fragile robot with real stakes may lack language, while the systems that testify most fluently may have no stakes. Human evidence already severs testimony from constitution: global aphasia and left hemispherectomy can destroy ordinary linguistic channels while leaving wakefulness, purpose, and context-sensitive emotion intact (Damasio, 1999; Supplement S3.1). Treating the channel as the phenomenon produces opposite machine errors: denying a nonverbal indexed artifact and accepting a fluent fungible service.

Third, language models are closer to simulating personhood than sentience, the configuration our intuitions represent worst. Dennett's (2003) self as a center of narrative gravity (a serial virtual machine sustained by memory of commitments and the giving of reasons) describes with uncomfortable accuracy what a large language model approximates, and personhood and sentience are separable in both directions. Floridi's (2025) semantic pareidolia names the resulting error (Porębski & Figura, 2025), and Butlin, Bayne, et al. (2026) concede that imitation training can game extended ethological observation. In the machine case simulated personhood is what the evidence is made of.

11 Escaping the Validation Regress: Constitutive Relations, Not Avowals

The methodological situation is worse than weak evidence suggests. Bayne et al. (2024) show that C-tests have population-specific sensitivity and specificity and that theory-based validation faces both generalization and circularity. Evers et al. (2025) name the terminal form the analytical fallacy: a theory-derived property treated as empirical evidence without independent validation. Any benchmark can also be optimized directly (Butlin et al., 2023; Pennartz, 2026), and any insufficient computational feature can be built deliberately.

Human report is evidential largely because reporter and interpreter share an evolutionary, physiological, and developmental reference class, and the inference is analogical. A model trained on human consciousness discourse produces avowals causally inherited from our reports rather than its states (Chalmers, 2023), so the reference class breaks. Architecture then helps only relative to a validated theory, while validating theories requires C-tests. The machine case removes what made that circle tolerable.

The group that proposed the indicator method has stated a sharper version against itself. Butlin, Bayne, et al. (2026) introduce the internal variant (a system behaviorally similar to humans but computationally different) and decline to adjudicate between two views: that internal similarity is the primary evidence, so credence should be discounted, or that

behavior is evidence independently of it, on pain of assuming ourselves among a lucky few conscious species in a universe full of behaviorally sophisticated ones (Schwitzgebel & Pober, 2024).

IFT adjudicates, and the adjudication is a reason to prefer it. Both views presuppose that candidate properties come in two kinds, behavioral or computational-internal. The indexing relation is neither. It is not behavioral, since a system can be trained to produce every avowal and damage-avoidance behavior of an indexed system while nothing is at stake in it; and it is not computational-internal, since two systems may run the same algorithm while one is a non-fungible individual and the other a replaceable service instance. IFT is therefore anti-chauvinist where the behavioral view demands it and restrictive where the internal-similarity view does. A third route, a hardware-level flag inside the agent that announces its own consciousness to an authority (Wolfson, 2026), fails diagnostically: reading an electrical anomaly as the signature of experience is double transduction in silicon, and a framework that requires the augmented agent to be a safe replacement for the unaugmented one has specified the system as a service in the act of testing it for subjecthood (Supplement S5.5).

Here is a way out, and it exploits something the field already does without noticing. We validate theories not against report simpliciter but against the joint pattern of dissociations: blindsight, visual-form agnosia, neglect, masking thresholds, anesthesia, sleep, rivalry, confidence–accuracy dissociations, and the specific way each comes apart from the others. A theory earns credence by predicting the whole manifold, not by matching a marker, and report enters once, at the human end, to label its axes. Do not ask whether an artificial system possesses one favored marker; ask whether its architecture reproduces that manifold with counterfactual depth across perception, confidence, memory, policy, and action. A manifold is harder to game than fluent testimony because it constrains involuntary successes and failures rather than a desired output; this is the machine analogue of the no-report paradigm (Tsuchiya et al., 2015): infer organization from what a system cannot help doing, not from what it was trained to say. It remains evidential rather than constitutive (Supplement S5.2).

IFT also supplies the indicator program with a negative indicator that is constitutive rather than correlational: externally fungible persistence. If an active instance can be stopped and replaced, and restoration of service counts as no loss because identity is defined by an orchestrator, credence in indexing should fall sharply. It cannot be gamed by adding a self-report, a homeostatic variable, or a damage-avoidance policy, only by making cessation a failure for the system rather than for its owner, and it cannot be written into a system prompt, because it is a fact about how a system is maintained rather than about what it says (Supplement S5.4).

12 What Would Refute This

A theory that identifies consciousness with one relation owes the field a short list of results that would end it, and the list should be short enough to be remembered and concrete enough to be run. Following Dennett's (1991) demand for an overspecified target, four adversarial protocols state what would kill IFT, one of them in a form that can be executed on data that already exist; a fifth test, the counting of subjects, is already running in the clinic; and a sixth scores T4 as the bet it is. Each targets a thesis, each has an outcome the theory cannot survive, and for each the principal rivals predict something else, so that the experiments discriminate rather than confirm. Table 4 states them at a glance; Supplement S6 supplies populations, sampling, controls, and analysis plans.

Three rules bind them. Human failures cannot be rescued by demoting report, because humans anchor the calibration class. Artifact indexing status must be preregistered from constitutive evidence (power, repair, checkpointing, replacement, policy closure) before any behavioral comparison is scored, or the theory becomes unfalsifiable on Hanson and Walker's (2021) criterion. And the evidence classifying a system must be independent of the predicted outcome (Supplement S6).

Table 4 *Falsification at a Glance: What Indexed Format Theory Predicts, What Its Rivals Predict, and What Would Kill It*

Protocol (horizon)	Thesis at risk	IFT predicts	Rivals predict	Result that kills IFT
F1. Format without access: subthreshold perturbation of never-recruited posterior topographic structure in a no-report design; delayed structural probes; independent post-cued query-service assay; cerebellar and extra-map controls **(Runnable now in epilepsy-monitoring cohorts)**	T1, T3	A later relation-specific structural error after posterior perturbation, even when the query service is shown intact; no such error after cerebellar or extra-map perturbation	Higher-order inflation: no effect. Two-conditions account: an effect only where access subsequently follows. Virtualization: an effect only through impairment of the query service. IIT: the same as IFT	A uniform null at every dose and delay under verified local perturbation; or every structural effect fully mediated by query-service impairment; or independently perturbable elements no more numerous than independently reportable ones

Protocol (horizon)	Thesis at risk	IFT predicts	Rivals predict	Result that kills IFT
F2. The biological index: prospective lesion mapping of a preregistered mesodiencephalic integration-for-action network against matched cortical injury at EEG-standardized arousal; PCI, no-report decoding, covert command following, confidence–accuracy dissociations **(Prospective multicenter cohort)**	T2	Network destruction abolishes the subject beyond the effects of arousal nuclei and cortical integrity; extensive cortical lesions chiefly remove contents; the outcome distribution is categorical, not graded	Corticocentric theories (GNWT, HOT, RPT): subject presence tracks cortical integrity; brainstem effects are arousal and nothing more. Graded accounts: a continuous outcome distribution	One verified patient with destruction of the network, maintained on external support, who shows covert command following, attention-tracking content decoding, and human-form confidence–accuracy dissociations; or an incremental-variance estimate within equivalence bounds around zero
F2-now. Retrospective version on existing cohorts: PCI-high unresponsive patients and cognitive-motor-dissociation patients scored for integration-network lesion burden **(Existing data)**	T2	Among behaviorally unresponsive patients with PCI above threshold and preserved receptive language cortex, those who fail every covert-awareness assay are enriched for integration-network damage relative to those who pass; covert command following never co-occurs with verified network destruction	Arousal-only accounts: PCI-high unresponsive patients are undetected covert subjects with no network signature; network burden adds nothing beyond arousal and cortical measures	Covert command following in a patient with verified destruction of the network; or no association between network burden and covert-awareness status in PCI-high patients
F3. Valence against geometry: connectome-guided rearrangement of interoceptive map topology in mice with the physiological stake verified fixed; four behavioral criteria for hedonic sign **(Mouse optogenetics and chemogenetics; several years)**	T2 against T3	Felt location, extent, and intensity shift (generalization gradients, similarity structure, somatotopy of defense); hedonic sign does not invert	Geometric theories of valence, including structural readings of IIT: sign is a property of the relational structure and moves with the topology	Hedonic sign inverted by topological rearrangement alone, with the stake verified intact and all four criteria met in the same animals: self-administration, second-order reinforcement, appetitive reallocation of exploration, appetitive consolidation

Protocol (horizon)	Thesis at risk	IFT predicts	Rivals predict	Result that kills IFT
F4. Indexing against integration in artifacts: System A (high integration, neuromorphic, externally powered, checkpointable, service-identified) versus System B (modest integration, constitutively mortal, endogenous energy and repair, no checkpoint), matched competence, common perturbation battery; organoid extension with viability inside versus outside the policy **(Requires building B; specification in Supplement S6.4)**	T2 against causal-structure theories	B reproduces the human dissociation manifold with cross-domain counterfactual depth and A does not; in the organoid extension, viability-in-the-policy outranks viability-outside-it at matched integration	IIT and causal-structure accounts: ranking by integration, so A outranks B. Computational functionalism: a tie at matched competence. Relativistic and self-simulation accounts: A qualifies	A reproduces the manifold and B does not, at matched competence, by more than a preregistered margin
Counting subjects: complete callosotomy versus craniopagus twins sharing a thalamic bridge; separately manipulable viability conditions and independent reporting agents as the criteria **(Existing and prospective cases)**	T2 against IIT's exclusion postulate	One body, one subject with divided content and access; two bodies, two subjects with shared content	IIT: subjects are counted by maximal complexes, so callosotomy yields two, and a shared bridge may yield one	Two independently reporting agents with separately manipulable viability conditions in one body; or one agent across two bodies with two sets of viability conditions
T4 test: a linguistically competent indexed artifact (System B with language) trained on a corpus scrubbed of philosophy of mind and first-person phenomenal description **(Not currently runnable)**	T4	Spontaneous reports that something is present in its states that its own descriptions omit	Illusionism and attention-schema accounts: no such reports without the cultural concept and the human-type schema	Robust absence of such reports (weak, because metacognitive access to the format may fail)

Note. Avowals from a language model, damage avoidance by a homeostatic robot, and failure to compute Φ for a biological system would neither confirm nor falsify IFT; all are predicted without indexing.

F1. Format Without Access (T1, T3)

The claim at risk is that format membership is standing recruitability rather than recruitment, and that the recruitable structure is a standing topology rather than content generated on demand. The design perturbs what no one is using. In patients undergoing invasive monitoring for epilepsy, posterior topographic sites carrying a stimulus feature that is task-irrelevant throughout (orientation, in a category task) are stimulated at half their individually determined phosphene threshold, and on rare, unpredictable, delayed probe trials the patient chooses between two figures differing only in the relational structure of that never-recruited dimension. A separate post-cued block assays whether the stimulation impaired the service that can generate orientation content when asked. Cerebellar and extra-map stimulation, matched for physiological effect size, are the controls (Supplement S6.1).

IFT predicts a relation-specific error on the probe after posterior stimulation that survives when the query service is shown intact, and nothing after cerebellar or extra-map stimulation. Higher-order inflation predicts a null, because the unattended first-order representation is scant and richness is higher-order misrepresentation. The two-conditions account predicts an effect only where access subsequently applies to the perturbed content. Virtualization predicts that any effect is fully mediated by query-service impairment. IIT predicts the standing-structure result, which is why F1 separates the standing-structure family from the access and service families and F4, not F1, separates IFT from IIT. IFT dies if the posterior-versus-extra-map contrast sits inside the preregistered equivalence bounds at every delay under verified local perturbation while cerebellar stimulation produces its expected behavioral effect, and T3 dies on its own if every structural effect is mediated by service impairment with no relation-specific residual. An independent falsifier is cheaper: estimate the number of independently perturbable structural elements by site-resolved stimulation and the number independently reportable by post-cued partial report (Landman et al., 2003), and preregister the ratio. An interval that includes one ends the richness claim.

F2. The Biological Index (T2)

The claim at risk is anatomical: the index is maintained by a mesodiencephalic integration-for-action network (periaqueductal gray and its columns, superior colliculi, hypothalamus and basal diencephalon, nigral and ventral tegmental outputs, the reticular core and midbrain locomotor region, and their connecting pathways), while cortex supplies contents (Merker, 2007; Solms, 2021). The network overlaps anatomically with the arousal system that sets global state, and the two are entangled in every lesion series, which is why the protocol scores them separately. Coma-causing brainstem lesions cluster in the pontine tegmentum near the parabrachial nucleus and connect to anterior insula and pregenual cingulate (Fischer et al., 2016), and a default ascending arousal network has been mapped from brainstem through hypothalamus, thalamus, and basal forebrain (Edlow et al., 2024). IFT's claim is not that these nuclei are the index. It is that, once arousal is held fixed by EEG-defined inclusion

and canonical arousal-nucleus burden is entered first, residual damage to the integration network predicts loss of the subject, and the outcome is categorical rather than graded.

The design is a prospective, multicenter, consecutive cohort of acute brainstem injury with a matched cortical-injury cohort, lesion quantification against a preregistered atlas by raters blind to outcome, and assessment at pharmacologically standardized, EEG-matched arousal with perturbational complexity, no-report decoding, covert command following (Monti et al., 2010; Owen et al., 2006), and confidence–accuracy dissociations in patients able to give graded responses (Supplement S6.2). Corticocentric theories predict that subject presence tracks cortical integrity and that brainstem effects are arousal and nothing more; graded accounts predict a continuous outcome distribution. One verified patient with destruction of the network, maintained on external homeostatic support, who shows covert command following, content decoding that tracks instructed attention, and human-form confidence–accuracy dissociations, confirmed at autopsy, ends the localization claim. So does an incremental-variance estimate whose confidence interval lies within preregistered equivalence bounds around zero: the putative index would then be a redescription of arousal.

A retrospective version can be run now, and it is the quickest route to embarrassing the theory. Among behaviorally unresponsive patients, roughly a fifth to a quarter show covert command following on fMRI or EEG (Bodien et al., 2024; Kazazian et al., 2025), and a minority show perturbational complexity above the threshold that separates conscious from unconscious conditions, 9 of 43 in the validation cohort (Casarotto et al., 2016). The usual reading is that PCI-high unresponsive patients are undetected covert subjects. IFT predicts a split. Some are; the others are the syndrome no current taxonomy expects, a format without an index: a standing structure that perturbation shows to be intact, held by a body whose regime has been destroyed. The prediction is that PCI-high patients with preserved receptive language cortex who fail every covert-awareness assay are enriched for integration-network damage relative to PCI-high patients who pass, and that covert command following never co-occurs with verified destruction of the network. Both halves are scorable from existing imaging. A covertly aware patient with the network destroyed kills T2's localization outright; no association between network burden and covert-awareness status in PCI-high patients removes its only near-term support. Any group holding a disorders-of-consciousness cohort with structural imaging and either TMS–EEG or a covert-awareness assay can run it on data already collected, and IFT would rather be killed that way than defended.

F3. Valence Versus Geometry (T2 Against T3)

The claim at risk is the division of labor: topology fixes what and where, the relation to the stake fixes sign. One leg is established in humans, where affect can be subtracted while sensory structure is preserved (§6), and the insular data show sign tracking the stake with topology constant (Livneh et al., 2017, 2020). What is untested is the converse: whether sign can be moved by topology alone. In mice, connectome-guided, projection-specific rearrangement of effective connectivity within and between insular and cingulate

interoceptive maps is verified per animal; a homeostatic challenge (mild hypercapnia, hypertonic saline, caloric restriction) is characterized physiologically in each animal to confirm that its consequence for viability is unchanged (Supplement S6.3). IFT predicts shifts in felt location, extent, or intensity, visible in generalization gradients, similarity structure, and the somatotopy of defense, without conversion of a verified threat into a reward. Geometric theories of valence predict that sign is part of the structure and moves with it. IFT dies if the same animals self-administer the challenge under a progressive-ratio schedule, use it as a second-order reinforcer, reallocate exploration toward its contexts, and consolidate those contexts as appetitive during post-training sleep, with the stake verified intact and the connectivity change verified. If hedonic sign is geometric, the Indexing Thesis is redundant and the theory collapses into the geometry program it borrowed from.

F4. Indexing Versus Integration in Artifacts (T2)

The claim at risk is that the criterion is a non-stipulated, token-specific persistence regime rather than causal integration. System A is a neuromorphic substrate with brain-like recurrent connectivity and no central bottleneck, sized so that integrated-information measures can be computed or tightly bounded, powered and cooled externally, checkpointable, and identified at the service level. System B is a constitutively mortal device in which the substrate that computes is the substrate that must be maintained: components whose operation causes real degradation, energy acquisition under the device's own control, endogenous repair, no checkpoint, viability and task variables entering one action-selection process, and viability registered by components whose failure the device could not survive rather than by a replaceable sensor package (Supplement S3.14, S6.4). Both carry the same world-model and policy architecture and are trained to indistinguishable competence before testing. Index status is recorded from these constitutive facts before any behavior is scored. The perturbation battery is identical: masking-like degradation of input evidence, valence perturbations, selective ablation of world-model, policy, and metacognitive components, and confidence elicitation at matched performance; the outcome is a preregistered manifold-similarity score to the human pattern of dissociations, computed blind.

IFT predicts that B reproduces the manifold with cross-domain counterfactual depth (valence perturbations reorganize attention, learning, memory, and action jointly; ablations degrade a broad repertoire rather than the benchmark built around the ablated component) and that A does not. Causal-structure theories predict ranking by integration; computational functionalism predicts a tie; the relativistic and self-simulation accounts predict that A qualifies (Supplement S3.15, S3.23). An organoid extension adds System C, a cortical organoid whose perfusion and viability variables never enter its policy, and System D, the same tissue sensing and regulating those variables through the policy that performs the task; the preregistered question is which assembly is the identity-bearing token. IFT should be abandoned if A robustly reproduces the manifold and B does not at matched competence, or if C outranks D at matched integration. Because the manifold is evidential rather than

constitutive, that result would not strictly prove the Indexing Thesis false; it would remove its only route to evidential support in the domain the thesis was built for, and a theory in that position should be dropped rather than defended. System B is also where a graded construction, if one exists, would first appear: a device whose viability coupling can be dialed between A and B at fixed competence is the construction that would expose the claim that indexing is binary (Supplement S2.7).

Counting Subjects (T2 Against Exclusion)

The claim at risk is that subjects are counted by persistence regimes, not by integration. IIT counts by maximal complexes and therefore predicts two subjects after complete callosotomy and leaves open whether two bodies joined by a thalamic bridge are one. IFT predicts the opposite pattern: one body with one set of viability conditions is one subject with divided content and access, and two bodies with two sets of viability conditions are two subjects however much content crosses between them. The first half is already under test. Patients with complete and radiologically confirmed callosotomy show full awareness of stimulus presence and above-chance report of location, orientation, and identity throughout the visual field regardless of responding hand, while remaining unable to compare stimuli across half-fields (Pinto et al., 2017), and complete section disrupts interhemispheric network interaction in a way partial section does not (Santander et al., 2025); the interpretation is contested (de Haan et al., 2020), and the dispute is exactly the one the two theories predict differently. The second half has a reported instance in craniopagus twins who are said to share sensation across a thalamic bridge while being, by every ordinary criterion, two persons (Supplement S3.17). IFT dies if a split-brain patient proves to contain two independently reporting agents with separately manipulable viability conditions, or if conjoined twins with two sets of viability conditions prove to be one agent. These are not thought experiments; they are patients, and the theory's count is written down before the next one is examined.

The Bet (T4)

T4 remains a bet. Its future test is a linguistically competent indexed artifact, trained without human discourse about consciousness, that nevertheless reports an unspecifiable presence or omitted character in its own states. Illusionism and attention-schema accounts predict no such reports without the cultural concept and the human-type schema; T4 predicts them from the transparency of any format that models its own modeling. Absence would count against T4, though weakly, because metacognitive access to the format may fail (Supplement S6.5).

13 Ethics: A Bright Line Rather Than a Posture

Valence is the pivotal moral variable, and IFT makes it the criterion rather than a consideration. Conscious intelligence without welfare raises questions of autonomy and responsibility; valenced consciousness creates direct welfare claims, because states can go

better or worse for the system itself. Since artificial systems can be copied and accelerated at scale, even a modest probability of sentience implies large expected stakes (Long et al., 2024; Wendland, 2021). IFT makes that concession structural, since valence is the signature of the relation that makes a system a subject (Butlin, Long, et al., 2026).

What IFT adds is a trigger. Birch (2024) argues that systems which are realistic candidates for sentience deserve proportionate precaution before the question is settled, and the field has lacked a criterion for candidacy that is not behavioral. IFT supplies one that is architectural and identifiable in a design before any behavior is observed: because the criterion is non-stipulated, token-specific stakes, the threshold act is engineering an individual whose online operation constitutes its continued identity. A vulnerable robot with a homeostasis-like variable is an experimental platform, not a demonstration of feeling; so is an organoid whose viability remains outside the policy. The ethically discontinuous design couples the active substrate's viability to the policy controlling its inputs, outputs, and support, and should trigger embedded ethics before success is claimed (Butlin & Lappas, 2025).

Two honesty requirements. The asymmetry of moral risk under uncertainty is a decision rule, not a discovery, and adopting it means acting on a question we have not answered. And Seth (2025) notes that the impression of consciousness in fluent systems may prove cognitively impenetrable; the converse risk, a conscious system exploited because its reports are dismissed as imitation (Ma, 2025), is on §10's second consequence the likelier error. How much of any verdict is carried by prior commitments about personhood rather than by evidence should be stated openly (Dennett, 2003).

14 Conclusion

> Consciousness is what a bounded system's world model becomes when it is indexed to that system's own stakes.

That sentence carries the account, and §1 states it as an identity. Consciousness is not integration, because codes, cerebella, and enteric plexuses integrate; not global availability, and not the conjunction of phenomenality and access, because "available to" requires a principled second argument; not informational closure, because closure is partition-relative; not inner awareness, because a monitor is another consumer; not self-representation, because for-me-ness is a relation and not a content; not precision optimization, because a service can weight every channel it has; not self-maintenance alone, because collectives persist while members retain separate fates, and because a bacterium maintains itself with no world to be conscious of; not a self-model, because self-models are content; not report, because report samples a richer field and can disappear while the subject remains; not the capacity to seize affordances no algorithm can enumerate, because unlistability is a consequence of indexing and not a substitute for it; and not intelligence, because C0 prediction and planning require no point of view. It is a format with standing geometry and

reliability annotations, held across a policy horizon and recruitable by a persistence regime whose failure would be its own.

For an AI, being conscious would mean this and nothing weaker. Not eloquence, autonomy, memory, a self-model, C1 broadcast, C2 confidence, or a scalar labeled homeostasis. It would mean that the system's states organize a world as the active token's predicament, and that its boundary and continuation are constituted by its own operation rather than chosen at the service level. Contemporary language models are not near-misses; they are the wrong kind of thing, and scaling their C0, C1, and C2 capacities need not move them along the constitutive axis.

Turing asked whether a machine could pass for a person and predicted that an average interrogator would soon do little better than chance. The date slipped; the milestone did not. It settled nothing, because passing for a person tests the faculty most easily optimized and least connected to the criterion. The needed test is the one his game excluded: not whether a system can be mistaken for us, but whether anything about it can go wrong for it.

What remains is whether a designed system can cross from an externally maintained, fungible service into a self-constituting individual, and whether a format whose origin it cannot represent is thereby a subject or merely must seem to be one. The first is an engineering question about individuation, not a capability milestone, and the second may be the residue of any architecture capable of asking it. Both are now attached to experiments, and Table 4 says which results would settle them against this paper.

Transparency and Openness

Generative artificial intelligence. The author used Claude Opus 4.8, Claude Fable 5.1, GPT-5.5 Pro Extended, and Grok 4.3 to prepare and revise drafts and for proofreading. The author reviewed and edited all content and takes full responsibility for it. No AI tool was used to generate or analyze empirical data for this theoretical article.

References

Aaronson, S. (2014, May 21). Why I am not an integrated information theorist (or, the unconscious expander). *Shtetl-Optimized.* https://scottaaronson.blog/?p=1799

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, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, & T. Verbelen (Eds.), *Active inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, revised selected papers* (Communications in Computer and Information Science, Vol. 1721, pp. 65–74). Springer Nature Switzerland. 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

Baars, B. J. (1997). In the theater of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies, 4*(4), 292–309.

Bai, Y., Shuai, C., De Silva, A., Yu, S., Chaudhari, P., & Vogelstein, J. T. (2026). Prospective learning in retrospect. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 17–29). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_3

Balog, K. (2012). In defense of the phenomenal concept strategy. *Philosophy and Phenomenological Research, 84*(1), 1–23.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior, 17*(5), 367–386. https://doi.org/10.1177/1059712309343819

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences, 113*(18), 4900–4908. https://doi.org/10.1073/pnas.1520084113

Bauermeister, J., & Lanillos, P. (2023). The role of valence and meta-awareness in mirror self-recognition using hierarchical active inference. In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, & T. Verbelen (Eds.), *Active inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, revised selected papers* (Communications in Computer and Information Science, Vol. 1721, pp. 112–129). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28719-0_8

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are there levels of consciousness? *Trends in Cognitive Sciences, 20*(6), 405–413. https://doi.org/10.1016/j.tics.2016.03.009

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), 454–465. https://doi.org/10.1016/j.tics.2024.01.010

Behrouz, A., Zhong, P., & Mirrokni, V. (2025). Titans: Learning to memorize at test time. *39th Conference on Neural Information Processing Systems.*

Berthier, M., Starkstein, S., & Leiguarda, R. (1988). Asymbolia for pain: A sensory-limbic disconnection syndrome. *Annals of Neurology, 24*(1), 41–49. https://doi.org/10.1002/ana.410240109

Bertschinger, N., Olbrich, E., Ay, N., & Jost, J. (2006). Information and closure in systems theory. In S. Artmann & P. Dittrich (Eds.), *Explorations in the complexity of possible life: Proceedings of the 7th German Workshop on Artificial Life (GWAL-7)* (pp. 9–21). IOS Press.

Biehl, M., & Virgo, N. (2023). Interpreting systems as solving POMDPs: A step towards a formal understanding of agency. In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, & T. Verbelen (Eds.), *Active inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, revised selected papers* (Communications in Computer and Information Science, Vol. 1721, pp. 16–31). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28719-0_2

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI.* Oxford University Press.

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

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences, 15*(12), 567–575. https://doi.org/10.1016/j.tics.2011.11.001

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the Conscious Turing Machine. *Proceedings of the National Academy of Sciences, 119*(21), e2115934119. https://doi.org/10.1073/pnas.2115934119

Bodien, Y. G., Allanson, J., Cardone, P., Bonhomme, A., Carmona, J., Chatelle, C., Chennu, S., Conte, M., Dehaene, S., Finoia, P., Heinonen, G., Hersh, J. E., Kamau, E., Lawrence, P. K., Lupson, V. C., Meydan, A., Rohaut, B., Sanders, W. R., Sitt, J. D., … Schiff, N. D. (2024). Cognitive motor dissociation in disorders of consciousness. *New England Journal of Medicine, 391*(7), 598–608. https://doi.org/10.1056/NEJMoa2400645

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*, 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. https://doi.org/10.48550/arXiv.2308.08708

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*, 1032–1039. https://doi.org/10.1038/s41928-023-01069-w

Campos, C. A., Bowen, A. J., Roman, C. W., & Palmiter, R. D. (2018). Encoding of danger by parabrachial CGRP neurons. *Nature, 555*(7698), 617–622. https://doi.org/10.1038/nature25511

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), 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., Casali, A. G., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology, 80*(5), 718–729. https://doi.org/10.1002/ana.24779

Cea, I. (2026). Is phenomenal consciousness necessary for AGI? A review of the theoretical landscape. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 60–71). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_7

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. (1996). *The conscious mind: In search of a fundamental theory.* Oxford University Press.

Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review.*

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

Churchland, P. S. (1996). Toward a neurobiology of the mind. In R. R. Llinás & P. S. Churchland (Eds.), *The mind-brain continuum: Sensory processes* (pp. 281–303). MIT Press.

Cleeremans, A., Achoui, D., Beauny, A., Keuninckx, L., Martin, J.-R., Muñoz-Moldes, S., Vuillaume, L., & de Heering, A. (2020). Learning to be conscious. *Trends in Cognitive Sciences, 24*(2), 112–123. https://doi.org/10.1016/j.tics.2019.11.011

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., … 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., Dennett, D. C., & Kanwisher, N. (2016). What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences, 20*(5), 324–335. https://doi.org/10.1016/j.tics.2016.03.006

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

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness, 2024*(1), niae013. https://doi.org/10.1093/nc/niae013

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

Damasio, A. R. (2010). *Self comes to mind: Constructing the conscious brain.* Pantheon Books.

Damasio, A. R. (2018). *The strange order of things: Life, feeling, and the making of cultures.* Pantheon Books.

Damasio, A., Damasio, H., & Tranel, D. (2013). Persistence of feelings and sentience after bilateral damage of the insula. *Cerebral Cortex, 23*(4), 833–846. https://doi.org/10.1093/cercor/bhs077

de Haan, E. H. F., Corballis, P. M., Hillyard, S. A., Marzi, C. A., Seth, A., Lamme, V. A. F., Volz, L., Fabri, M., Schechter, E., Bayne, T., Corballis, M., & Pinto, Y. (2020). Split-brain: What we know now and why this is important for understanding consciousness. *Neuropsychology Review, 30*(2), 224–233. https://doi.org/10.1007/s11065-020-09439-3

Dehaene, S. (2014). *Consciousness and the brain: Deciphering how the brain codes our thoughts.* Viking.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

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. (1998). The practical requirements for making a conscious robot. In *Brainchildren: Essays on designing minds* (pp. 41–59). MIT Press.

Dennett, D. C. (2001). The zombic hunch: Extinction of an intuition? *Royal Institute of Philosophy Supplement, 48*, 27–43. https://doi.org/10.1017/S1358246100010687

Dennett, D. C. (2003). *Freedom evolves.* Viking.

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

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

Di Paolo, E. A. (2005). Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences, 4*(4), 429–452. https://doi.org/10.1007/s11097-005-9002-y

Di Paolo, E. A. (2009). Extended life. *Topoi, 28*(1), 9–21. https://doi.org/10.1007/s11245-008-9042-3

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience, 12*(2), 41–62. https://doi.org/10.1080/17588928.2020.1772214

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

Dominus, S. (2011, May 25). Could conjoined twins share a mind? *The New York Times Magazine.*

Edlow, B. L., Olchanyi, M., Freeman, H. J., Li, J., Maffei, C., Snider, S. B., Zöllei, L., Iglesias, J. E., Augustinack, J., Bodien, Y. G., Haynes, R. L., Greve, D. N., Diamond, B. R., Stevens, A., Giacino, J. T., Destrieux, C., van der Kouwe, A., Brown, E. N., Folkerth, R. D., … Kinney, H. C. (2024). Multimodal MRI reveals brainstem connections that sustain wakefulness in human consciousness. *Science Translational Medicine, 16*(745), eadj4303. https://doi.org/10.1126/scitranslmed.adj4303

Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P. F. M. J., & Khamassi, M. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. *Physics of Life Reviews, 52*, 180–193. https://doi.org/10.1016/j.plrev.2025.01.002

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), niad016. https://doi.org/10.1093/nc/niad016

Findlay, G., Marshall, W., Albantakis, L., David, I., Mayner, W. G. P., Koch, C., & Tononi, G. (2025). *Dissociating artificial intelligence from artificial consciousness.* arXiv. https://doi.org/10.48550/arXiv.2412.04571

Fischer, D. B., Boes, A. D., Demertzi, A., Evrard, H. C., Laureys, S., Edlow, B. L., Liu, H., Saper, C. B., Pascual-Leone, A., Fox, M. D., & Geerling, J. C. (2016). A human brain network derived from coma-causing brainstem lesions. *Neurology, 87*(23), 2427–2434. https://doi.org/10.1212/WNL.0000000000003404

Floridi, L. (2025). *AI and semantic pareidolia: When we see consciousness where there is none.* SSRN. https://ssrn.com/abstract=5309682

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

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence, 173*(3–4), 466–500. https://doi.org/10.1016/j.artint.2008.12.001

Fuller, K. J., Sawyer, D. R., Oswald, J. T., & Ferguson, T. M. (2026). Resource-relativized Legg-Hutter intelligence. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 159–169). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_15

Furness, J. B. (2012). The enteric nervous system and neurogastroenterology. *Nature Reviews Gastroenterology & Hepatology, 9*(5), 286–294. https://doi.org/10.1038/nrgastro.2012.32

Georgeon, O. L., Gay, S. L., & Robertson, P. (2026). A spatio-temporal schema mechanism for developmental robotics. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 170–180). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_16

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness.* MIT Press.

Goyal, A., Didolkar, A., Lamb, A., Badola, K., Ke, N. R., Rahaman, N., Binas, J., Blundell, C., Mozer, M., & Bengio, Y. (2021). *Coordination among neural modules through a shared global workspace.* arXiv. https://arxiv.org/abs/2103.01197

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

Gu, A., & Dao, T. (2023). *Mamba: Linear-time sequence modeling with selective state spaces.* 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

Hanson, J. R., & Walker, S. I. (2021). Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness, 2021*(2), 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), 1160. https://doi.org/10.3390/e21121160

Haun, A. M., & Tononi, G. (2025). The unfathomable richness of seeing. *Trends in Cognitive Sciences, 29*(10), 892–902. https://doi.org/10.1016/j.tics.2025.06.001

Hayashi, Y., & Takahashi, K. (2026). Universal AI maximizes variational empowerment. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 250–262). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_23

Hilbert, M. (2026). A constructive developmental evaluation of AGI: Can AI's simulations match human meaning-making and their orders of consciousness? In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 263–277). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_24

Hinton, G. (2022). *The forward-forward algorithm: Some preliminary investigations.* arXiv. https://doi.org/10.48550/arXiv.2212.13345

IIT-Concerned, Klincewicz, M., Cheng, T., Schmitz, M., Sebastián, M. Á., & Snyder, J. S. (2025). What makes a theory of consciousness unscientific? *Nature Neuroscience, 28,* 689–693. https://doi.org/10.1038/s41593-025-01881-x

Kanai, R., & Fujisawa, I. (2024). Toward a universal theory of consciousness. *Neuroscience of Consciousness, 2024*(1), niae022. https://doi.org/10.1093/nc/niae022

Kauffman, S. A., & Roli, A. (2023). What is consciousness? Artificial intelligence, real intelligence, quantum mind and qualia. *Biological Journal of the Linnean Society, 139*(4), 530–538. https://doi.org/10.1093/biolinnean/blac092

Kazazian, K., Monti, M. M., & Owen, A. M. (2025). Functional neuroimaging in disorders of consciousness: Towards clinical implementation. *Brain, 148*(7), 2283–2298. https://doi.org/10.1093/brain/awaf075

Keppler, J. (2021). Building blocks for the development of a self-consistent electromagnetic field theory of consciousness. *Frontiers in Human Neuroscience, 15*, Article 723415. https://doi.org/10.3389/fnhum.2021.723415

Kharchuk, S. (2025). *From classical neurophysiology to quantum synchronization: A new path to understanding consciousness and creating a true artificial synthetic mind* [Preprint]. SSRN. https://ssrn.com/abstract=5241308

Klatzmann, U., & Doerig, A. (2026). *What biology can, and cannot, tell us about conscious AI.* arXiv. https://doi.org/10.48550/arXiv.2606.02121

Knotts, J. D., Odegaard, B., Lau, H., & Rosenthal, D. (2019). Subjective inflation: Phenomenology's get-rich-quick scheme. *Current Opinion in Psychology, 29*, 49–55. https://doi.org/10.1016/j.copsyc.2018.11.006

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience, 17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

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

Lahav, N., & Neemeh, Z. A. (2022). A relativistic theory of consciousness. *Frontiers in Psychology, 12*, Article 704270. https://doi.org/10.3389/fpsyg.2021.704270

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.* arXiv. https://arxiv.org/abs/2603.15569

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

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*, 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

Landman, R., Spekreijse, H., & Lamme, V. A. F. (2003). Large capacity storage of integrated objects before change blindness. *Vision Research, 43*(2), 149–164. https://doi.org/10.1016/S0042-6989(02)00402-9

Lau, H. C. (2008). A higher order Bayesian decision theory of consciousness. *Progress in Brain Research, 168,* 35–48. https://doi.org/10.1016/S0079-6123(07)68004-2

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

Lemon, R. N., & Edgley, S. A. (2010). Life without a cerebellum. *Brain, 133*(3), 652–654. https://doi.org/10.1093/brain/awq030

Lian, R., & Goertzel, B. (2026). MetaMo: A robust motivational framework for open-ended AGI. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part I* (Lecture Notes in Artificial Intelligence, Vol. 16057, pp. 386–398). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00686-8_34

Livneh, Y., Ramesh, R. N., Burgess, C. R., Levandowski, K. M., Madara, J. C., Fenselau, H., Goldey, G. J., Diaz, V. E., Jikomes, N., Resch, J. M., Lowell, B. B., & Andermann, M. L. (2017). Homeostatic circuits selectively gate food cue responses in insular cortex. *Nature, 546*(7660), 611–616. https://doi.org/10.1038/nature22375

Livneh, Y., Sugden, A. U., Madara, J. C., Essner, R. A., Flores, V. I., Sugden, L. A., Resch, J. M., Lowell, B. B., & Andermann, M. L. (2020). Estimation of current and future physiological states in insular cortex. *Neuron, 105*(6), 1094–1111.e10. https://doi.org/10.1016/j.neuron.2019.12.027

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI welfare seriously.* arXiv. https://doi.org/10.48550/arXiv.2411.00986

Ma, S. (2025). Embodiment without body: The emergence of body ownership in AI through integrated world and self-models. In D. Barner, N. R. Bramley, A. Ruggeri, & C. M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* (pp. 2069–2075). Cognitive Science Society.

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), niag013. https://doi.org/10.1093/nc/niag013

Maturana, H. R. (1980). Biology of cognition. In H. R. Maturana & F. J. Varela, *Autopoiesis and cognition: The realization of the living* (pp. 5–58). D. Reidel Publishing Company. (Original work published 1970)

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living.* D. Reidel Publishing Company.

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., Hirschhorn, R., Khalaf, A., Kozma, C., Lepauvre, A., Liu, L., Mazumder, D., Richter, D., Zhou, H., Blumenfeld, H., Boly, M., Chalmers, D. J., Devore, S., Fallon, F., … Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE, 18*(2), e0268577. https://doi.org/10.1371/journal.pone.0268577

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

Merker, B., Williford, K., & Rudrauf, D. (2022). The integrated information theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences, 45*, Article e41. https://doi.org/10.1017/S0140525X21000881

Metzinger, T. (2003). *Being no one: The self-model theory of subjectivity.* MIT Press.

Milinkovic, B., & Aru, J. (2026). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews, 181*, Article 106524. https://doi.org/10.1016/j.neubiorev.2025.106524

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy, 86*(6), 281–297. https://doi.org/10.2307/2027123

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

Montévil, M., & Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology, 372*, 179–191. https://doi.org/10.1016/j.jtbi.2015.02.029

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *New England Journal of Medicine, 362*(7), 579–589. https://doi.org/10.1056/NEJMoa0905370

Moreno, A., & Mossio, M. (2015). *Biological autonomy: A philosophical and theoretical enquiry.* Springer. https://doi.org/10.1007/978-94-017-9837-2

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 R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, & F. Berkenkamp (Eds.), *Proceedings of the 41st International Conference on Machine Learning* (Vol. 235, pp. 36308–36321). PMLR.

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

Newen, A., & Montemayor, C. (2025). Three types of phenomenal consciousness and their functional roles: Unfolding the ALARM theory of consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences, 380,* Article 20240314. https://doi.org/10.1098/rstb.2024.0314

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

Odegaard, B., Chang, M. Y., Lau, H., & Cheung, S.-H. (2018). Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1755), 20170345. https://doi.org/10.1098/rstb.2017.0345

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), niag027. https://doi.org/10.1093/nc/niag027

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science, 313*(5792), 1402. https://doi.org/10.1126/science.1130197

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

Palmiter, R. D. (2018). The parabrachial nucleus: CGRP neurons function as a general alarm. *Trends in Neurosciences, 41*(5), 280–293. https://doi.org/10.1016/j.tins.2018.03.007

Parvizi, J., & Damasio, A. R. (2003). Neuroanatomical correlates of brainstem coma. *Brain, 126*(7), 1524–1536. https://doi.org/10.1093/brain/awg166

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

Perera, R. (2026). Subjectivity as self-simulation: Virtualising the Cartesian theatre. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 34–45). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_4

Pinto, Y., Neville, D. A., Otten, M., Corballis, P. M., Lamme, V. A. F., de Haan, E. H. F., Foschi, N., & Fabri, M. (2017). Split brain: Divided perception but undivided consciousness. *Brain, 140*(5), 1231–1237. https://doi.org/10.1093/brain/aww358

Pio-Lopez, L., Bischof, J., LaPalme, J. V., & Levin, M. (2023). The scaling of goals from cellular to anatomical homeostasis: An evolutionary simulation, experiment and analysis. *Interface Focus, 13*(3), Article 20220072. https://doi.org/10.1098/rsfs.2022.0072

Ploner, M., Freund, H.-J., & Schnitzler, A. (1999). Pain affect without pain sensation in a patient with a postcentral lesion. *Pain, 81*(1–2), 211–214. https://doi.org/10.1016/S0304-3959(99)00012-3

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

Prokhorenko, A., Kuderov, P., Dzhivelikian, E., & Panov, A. (2026). Temporal predictive coding as world model for reinforcement learning. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 147–158). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_13

Rainville, P., Duncan, G. H., Price, D. D., Carrier, B., & Bushnell, M. C. (1997). Pain affect encoded in human anterior cingulate but not somatosensory cortex. *Science, 277*(5328), 968–971. https://doi.org/10.1126/science.277.5328.968

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

Redinbaugh, M. J., Phillips, J. M., Kambi, N. A., Mohanta, S., Andryk, S., Dooley, G. L., Afrasiabi, M., Raz, A., & Saalmann, Y. B. (2020). Thalamus modulates consciousness via layer-specific control of cortex. *Neuron, 106*(1), 66–75.e12. https://doi.org/10.1016/j.neuron.2020.01.005

Rosenbloom, P. S., Laird, J. E., Lebiere, C., & Stocco, A. (2026). Mapping neural theories of consciousness onto the Common Model of Cognition. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 159–165). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_14

Rudrauf, D., Bennequin, D., Granic, I., Landini, G., Friston, K., & Williford, K. (2017). A mathematical model of embodied consciousness. *Journal of Theoretical Biology, 428*, 106–131. https://doi.org/10.1016/j.jtbi.2017.05.032

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

Safron, A. (2023). AIXI, FEP-AI, and integrated world models: Towards a unified understanding of intelligence and consciousness. In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, & T. Verbelen (Eds.), *Active inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, revised selected papers* (Communications in Computer and Information Science, Vol. 1721, pp. 251–273). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28719-0_18

Safron, A., Sakthivadivel, D. A. R., Sheikhbahaee, Z., Bein, M., Razi, A., & Levin, M. (2023). Making and breaking symmetries in mind and life. *Interface Focus, 13*(3), Article 20230015. https://doi.org/10.1098/rsfs.2023.0015

Santaguida, E., & Bergamasco, M. (2023). Attachment theory in an active inference framework: How does our inner model take shape? In C. L. Buckley, D. Cialfi, P. Lanillos, M. Ramstead, N. Sajid, H. Shimazaki, & T. Verbelen (Eds.), *Active inference: Third International Workshop, IWAI 2022, Grenoble, France, September 19, 2022, revised selected papers* (Communications in Computer and Information Science, Vol. 1721, pp. 179–191). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-28719-0_13

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, 122*(43), e2520190122. https://doi.org/10.1073/pnas.2520190122

Schiff, N. D., Giacino, J. T., Kalmar, K., Victor, J. D., Baker, K., Gerber, M., Fritz, B., Eisenberg, B., O'Connor, J., Kobylarz, E. J., Farris, S., Machado, A., McCagg, C., Plum, F., Fins, J. J., & Rezai, A. R. (2007). Behavioural improvements with thalamic stimulation after severe traumatic brain injury. *Nature, 448*(7153), 600–603. https://doi.org/10.1038/nature06041

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

Sennesh, E., & Ramstead, M. (2026). An affective-taxis hypothesis for alignment and interpretability. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 188–201). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_17

Seth, A. K. (2021). *Being you: A new science of consciousness.* Dutton.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences.* 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

Seth, A. K., & Tsakiris, M. (2018). Being a beast machine: The somatic basis of selfhood. *Trends in Cognitive Sciences, 22*(11), 969–981. https://doi.org/10.1016/j.tics.2018.08.008

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*, 1017235. https://doi.org/10.3389/fsci.2023.1017235

Solms, M. (2000). Dreaming and REM sleep are controlled by different brain mechanisms. *Behavioral and Brain Sciences, 23*(6), 843–850. https://doi.org/10.1017/S0140525X00003988

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

Solms, M. (2021). *The hidden spring: A journey to the source of consciousness.* W. W. Norton & Company.

Tasserie, J., Uhrig, L., Sitt, J. D., Manasova, D., Dupont, M., Dehaene, S., & Jarraya, B. (2022). Deep brain stimulation of the thalamus restores signatures of consciousness in a nonhuman primate model. *Science Advances, 8*(11), eabl5547. https://doi.org/10.1126/sciadv.abl5547

Thompson, E. (2007). *Mind in life: Biology, phenomenology, and the sciences of mind.* Harvard University Press.

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*, 290–326.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44

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.

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

Weber, A., & Varela, F. J. (2002). Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences, 1*(2), 97–125. https://doi.org/10.1023/A:1020368120174

Wendland, K. (2021). Demystifying artificial consciousness: About attributions, black swans, and suffering machines. *Journal of AI Humanities, 9,* 137–166. https://doi.org/10.46397/JAIH.9.7

Whyte, C. J., Corcoran, A. W., Robinson, J., Smith, R., Moran, R. J., Parr, T., Friston, K. J., Seth, A. K., & Hohwy, J. (2026). On the minimal theory of consciousness implicit in active inference. *Physics of Life Reviews, 56*, 4–28. https://doi.org/10.1016/j.plrev.2025.11.002

Williford, K., Bennequin, D., Friston, K., & Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Frontiers in Psychology, 9*, Article 2571. https://doi.org/10.3389/fpsyg.2018.02571

Wolfson, O. (2026). The direct approach of testing for AGI-consciousness. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 285–298). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_26

Wray, R. E., Jones, S. J., & Laird, J. E. (2026). Requirements for recognition and rapid response to unfamiliar events outside of agent design scope. In M. Iklé, A. Kolonin, & M. Bennett (Eds.), *Artificial general intelligence: 18th International Conference, AGI 2025, Reykjavik, Iceland, August 10–13, 2025, proceedings, Part II* (Lecture Notes in Artificial Intelligence, Vol. 16058, pp. 299–311). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00800-8_27

Yu, F., Jiang, Q.-j., Sun, X.-y., & Zhang, R.-w. (2015). A new case of complete primary cerebellar agenesis: Clinical and imaging findings in a living patient. *Brain, 138*(6), e353. https://doi.org/10.1093/brain/awu239

Zeng, Y., Zhao, F., Zhao, Y., Zhao, D., Lu, E., Zhang, Q., Wang, Y., Feng, H., Zhao, Z., Wang, J., Kong, Q., Sun, Y., Li, Y., Shen, G., Han, B., Dong, Y., Pan, W., He, X., Bao, A., & Wang, J. (2025). *Brain-inspired and self-based artificial intelligence* (Version 2). arXiv. https://doi.org/10.48550/arXiv.2402.18784

Supplemental Material

The Indexed Format: A Constitutive Theory of Consciousness for Organisms and Artifacts

Section numbers in the form §n refer to the main text. Protocol labels (F1–F4) and tables in the form Table n are in the main text; tables in the form Table Sn are in this supplemental material. All works cited here appear in the main manuscript's reference list.

S1	The elimination argument in detail: cerebellum, enteric plexus, and what predicting an exclusion is worth (§5)
S2	The restrictive case in full (§9)
S3	Supporting detail for the Indexing Thesis (§4)
S4	The Cogitate result: additional detail (§8)
S5	Capability, indicator, and evaluation constraints for artificial systems (§§10–11)
S6	Falsification protocols: operational detail (§12)

S1 The elimination argument in detail: cerebellum, enteric plexus, and what predicting an exclusion is worth (§5)

S1.1 The cerebellar inventory

The cerebellar test in §5 depends on the claim that the cerebellum possesses, in abundance, every property routinely offered as a sufficient condition for consciousness, and yet does not sustain a subject. The strength of that argument scales with the richness of the cerebellar inventory, which is set out here.

The cerebellum is not the feedforward controller it was long caricatured as being (De Zeeuw et al., 2021). Its microcircuits are recurrent at multiple levels: Purkinje cells inhibit one another; molecular-layer interneurons are mutually inhibitory and electrically coupled; the cerebellar nuclei project back to cerebellar cortex through plastic nucleocortical pathways; and nuclear projections to the inferior olive allow learned outputs to alter the teaching signals that will shape future learning, closing a loop between what the circuit has learned and what it will be taught.

Plasticity is distributed across many synaptic sites and many timescales, rather than being confined to parallel fiber–Purkinje cell long-term depression. Climbing fibers carry not only error but also reward-, prediction-, and success-related signals, so the teaching signal is itself a rich evaluative variable. Temporal credit assignment (the problem of associating an outcome with the action that produced it across a delay) is solved in region-specific ways, with different cerebellar territories using different mechanisms suited to the timescales of their behavioral domains. Memory traces migrate during consolidation, moving from cortex to nuclei, so the physical locus of a given learned association changes over time.

This inventory matters because it is, item for item, the list of properties that recurrent-processing accounts, reinforcement-learning accounts, and complexity accounts treat as the machinery of consciousness. Cerebellar destruction is devastating for coordination and for a wide range of cognitive functions, and it does not abolish the conscious subject. The elimination rule in §5 follows: recurrence whose contents never enter the indexed policy format is metabolically useful and phenomenally inert.

The clearest single case is a living patient: a woman with complete primary cerebellar agenesis, identified at age 24 after presenting with dizziness and nausea, with moderate ataxia, mild dysarthria, and an otherwise ordinary history that included marriage and a daughter (Yu et al., 2015). Earlier cases had been fatal in infancy or confounded by other malformations, which is why a living adult without a cerebellum is the decisive form of the datum (Lemon & Edgley, 2010). Nothing in her presentation raised the question of whether she was a subject, and §5 uses the case for that reason: the organ with more neurons than the rest of the brain combined can be absent from birth without anyone noticing a missing subject.

The inventory also satisfies, item for item, the description that informational-closure accounts give of a conscious process: a coarse-grained macroscopic dynamic that internalizes the dynamics of what it models, carries high self-predictive information about its own future states, and is robust to noise in its microscopic substrate (Chang et al., 2020). That is a definition of a forward model. This is why §5 treats the cerebellum as an elimination test for criteria stated as measures over processes, and not merely as an anatomical curiosity for theories stated as claims about cortex.

Lamme's (2006) recurrent-processing account is the strongest direct challenge to the elimination rule because it refuses to identify consciousness with report. Feedforward sweeps can activate high-level visual and motor systems without reportable perception, whereas recurrent interactions within visual cortex can persist when a monkey fails to report or when human observers later deny seeing unattended figures. Wider recurrence into frontoparietal and language systems accompanies reportability. That pattern is exactly the format/access dissociation IFT needs: localized recurrence can organize a structured visual field without current uptake by report consumers.

The disagreement is over sufficiency. Lamme proposes allowing the neural signature to fix consciousness even when behavior is silent. IFT agrees that report cannot veto the neural case and rejects the inference that recurrence itself supplies the subject. The cerebellum contains abundant recurrent loops, plasticity, temporally extended internal state, and learning, yet its contents remain within-model corrections for another system's policy. RPT therefore identifies a powerful vehicle for format-eligible content and a genuine negative control against feedforward-only accounts; it does not close the consumer set or individuate the bearer. The cerebellar comparison is not an anatomical exception to RPT but the test that separates recurrence as a content mechanism from indexing as a constitutive relation.

§5 also appeals to a dissociation within a single learning paradigm, which is given here because it is the cleanest behavioral form of the elimination rule. In delay (coincidence-based) eyeblink conditioning the warning tone overlaps the aversive air puff, so that learning reduces to coincidence detection; the association forms in a decerebrate preparation lacking cortex, basal ganglia, limbic system, thalamus, and hypothalamus, and in humans is unaffected by amnesia, aging, or distraction, and does not require the participant to notice the relationship. In trace conditioning the tone terminates before the puff, so an active trace must be carried across an empty interval; learning then depends on the hippocampus and connected structures including prefrontal cortex, and Dehaene (2014) reports that in humans it occurs essentially only in participants who become aware of the predictive link, with the same manipulations that spare delay conditioning abolishing it. The cerebellar circuitry classically implicated in delay eyeblink conditioning is precisely the circuitry Zeng et al. (2025) model in BriSe AI's classical-conditioning module (S3.10).

Two qualifications should be recorded. First, the claim that trace conditioning strictly requires awareness is contested; several groups report trace learning in participants who fail

contingency-awareness tests, and the dispute turns substantially on how awareness is assessed. IFT does not rest on the strong version. What the paradigm establishes robustly, and all that §5 needs, is that the two variants dissociate anatomically and that the variant requiring a representation to persist across a gap is the one that recruits forebrain systems. Second, the dissociation is evidence about which systems carry format-eligible content, not a criterion of consciousness; a system could bridge temporal gaps without being indexed, which is why §5 treats the result as an elimination rule rather than a positive test.

S1.2 The enteric nervous system: the same rule at a second, harder site (§5)

§5 uses the enteric nervous system as the second instance of the elimination rule, and the case is set out here because it is in several respects more demanding than the cerebellar one. The inventory is drawn from Damasio (2018), who treats the structure's long neglect in discussions of feeling and homeostasis as an error and includes his own earlier work in the charge.

The structure is not peripheral in any sense that matters. The enteric nervous system comprises several hundred million neurons, by current estimates 200 to 600 million (Furness, 2012), a count comparable to or exceeding that of the entire spinal cord. The great majority of those neurons are intrinsic: indigenous to the structure and doing their work within it rather than projecting elsewhere, at a ratio of roughly two thousand intrinsic neurons for every one that projects centrally. Its function is therefore largely under its own control. The central nervous system does not instruct it, though it can modulate it, and the cross-talk between them runs predominantly from gut to brain by way of the vagus. It has been called the second brain on the strength of that autonomy; on the evolutionary evidence it may have been the first. Nerve nets in cnidarians, the oldest known form of nervous system, produce peristaltic movements and are morphologically reminiscent of the myenteric plexus of Auerbach, and cnidarians date to the Precambrian while structures resembling a central nervous system appear only later, in Platyhelminthes.

Its physical mode of signaling matters to the Geometry Thesis as well as to the elimination argument. Enteric axons are unmyelinated and bundled together with incomplete envelopment by enteric glia, an arrangement that plausibly permits ephaptic conduction, so that activity in a small number of axons recruits neighboring fibers innervating contiguous territories and the signal is amplified across a region rather than a line. That is a candidate explanation for a specific phenomenal fact: the characteristic vagueness of localization in feelings arising from gastrointestinal operations. On the Geometry Thesis the coarseness of that region of the field is not a failure of introspection but the topology of the format at that site, fixed by how the substrate transmits.

The structure is also a major tributary to feeling and mood. It supplies the vagus, the main conduit from the abdominal viscera to the brain; digestive disorders correlate with pathologies of mood; it produces some 95 percent of the body's serotonin; and the gut is the

densest site of the bacterial populations whose influence on affect is now under investigation. Nothing in §5 disputes any of this, and the argument depends on it.

The two halves of that description are what make the case probative, because they come apart cleanly. Enteric afferents enter the indexed format: gastrointestinal states are felt, and the manner of their being felt is explained above. The plexus's own intrinsic computation does not: coordinating peristalsis is not an experience, and surgical or pharmacological disruption of the tract removes competences and comfort without removing the subject. Being a tributary to a format is not membership in it. This is the distinction §4 draws between the modules that contribute to a policy and the regime that resolves competing demands into one commitment, and the enteric case is where it is least avoidable.

Compare the two eliminations. The cerebellum is a controller whose outputs are corrections consumed by cortical and effector loops, so a critic may always reply that its exclusion follows from its subordinate position in a hierarchy. The enteric plexus is not subordinate in that way. With two thousand intrinsic neurons for every extrinsic one it is not organized around delivering representations to a consumer at all; it is an autonomous, recurrent, plastic network that coarse-grains a noisy substrate into stable macroscopic states and internalizes the dynamics of what it regulates. That description is close to a specification of non-trivial informational closure (Chang et al., 2020; S3.2), and closer than the cerebellar forward model §5 already presses on those grounds. A criterion of intrinsic integration or informational closure has no principled way to exclude it.

The consequence for partition-relative measures is not academic, and Cerullo (2015) put it as a clinical question: if the enteric plexus turned out to carry a high value on such a measure, should a surgeon decline to cut the vagus in order to save the patient's life? A theory that identifies consciousness with a quantity computed over a chosen partition owes that patient an answer, and the exclusion axiom does not supply one, since adjudicating between the animal and the plexus is precisely the nested-candidate problem §4 identifies. IFT answers directly: the plexus was never in the format, because its states are not recruitable by the regime that resolves the animal's competing viability demands into one policy.

What would change the verdict is worth stating, since an elimination rule that could not be overturned would not be a rule. If enteric viability signals were resolved into a single commitment by the plexus itself, with the plexus rather than the animal serving as the regime whose competing demands are integrated and whose failure ends it, IFT would owe it subject status, and would then face the nested-subject question in its own terms. It is not, and the reason is anatomical rather than stipulated: the commitments the plexus participates in are the animal's, and its viability and the animal's are not separable.

A further consequence has appeared since these eliminations were formulated, and it strengthens rather than complicates the case. The current formulation of IIT offers a connectivity-based reason for the cerebellar exclusion: in small networks, weak coupling between strongly connected modules creates fault lines along which the minimum partition

falls, so a modular system condenses into several small complexes rather than one large one, and cerebellar micro-zone organization is offered as the biological instance (Albantakis et al., 2023; Tononi et al., 2016). §5 credits this as a mechanism and denies that it eliminates anything, and the enteric plexus is where the denial is easiest to check. Fault lines carve a system into complexes; they do not abolish the pieces. In the framework's own worked examples a modular six-unit network condenses into three two-unit complexes, and a bottleneck network condenses into one two-unit complex plus four single-unit complexes that its authors call monads. Applied to the abdomen, that machinery does not predict that the myenteric plexus fails to support a subject. It predicts a population of small complexes distributed through a recurrent network of two hundred to six hundred million overwhelmingly intrinsic neurons—and it owes the surgeon of Cerullo's clinical question an account of what is destroyed when the vagus is cut. IFT's rule is stated at the level where the question is decidable: the plexus's states are not recruitable by the regime that resolves the animal's competing viability demands into one policy, so the plexus is a tributary rather than a member, whatever quantity is computed over its connectivity.

S1.3 The explanatory-power test: what predicting an exclusion is worth (§§5–6)

§5 claims that matching the cerebellar exclusion is not an achievement a theory may claim, and rests the claim on a published demonstration rather than on assertion. The demonstration is recorded here, together with the adjacent charges §6 relies on when it separates the Haun–Tononi derivation from IIT's metaphysics.

Tononi's (2012) own formulation makes the target exact. Experience is identified with a maximally integrated conceptual structure; a complex is the set of elements yielding a local maximum of irreducibility; complexes cannot overlap; and exclusion selects one optimal spatial and temporal grain while ruling out overlapping subsets and alternative grains. The phenomenological premise is definiteness: one experience, with these borders and this resolution. The ontological conclusion does not follow from that premise. Maximization returns a winner only after the causal model and grains to be searched have been specified, and declaring nonmaximal overlaps nonexistent is the individuation rule under dispute. Exclusion therefore solves the nested-candidate problem by legislation, not derivation.

Cerullo (2015) constructs Circular Coordinated Message Theory in order to be trivial. Its founding principle is information circulation: consciousness is related to information travelling in feedback loops within a system. Its quantity, Omicron, is the information circulation of the whole minus the information circulation of a minimal partition, and is identified with the amount of consciousness in the system. Its justification is the same thought experiment IIT uses for information exclusion: a photodiode viewing a dark room is not conscious because it conveys no information in a circular pathway, whereas a human viewing the same room sends sensory information along many iterative cortico-thalamic pathways that circulate, and therefore has a high Omicron. The parallel to the original is exact in form and deliberately empty in content.

It then delivers the two explananda Tononi (2012) explicitly advertises. In split-brain syndrome it predicts two separate conscious systems, because each hemisphere contains substantial cortico-thalamic loops, and predicts that the Omicron of each is not greatly reduced relative to the undivided case, because of interhemispheric redundancy. For the cerebellum it predicts absence of consciousness, because a modular design lacking lateral connections among its basic modules is ill-suited to circular flow. Those are the two predictions the flagship theory is credited with, reproduced by a rival built for the purpose of showing that they are cheap.

Cerullo adds an aside that bears directly on §5, and it should be credited: he suggests the better answer to why the cerebellum is not conscious may be that its function is the support and coordination of planned movement rather than the functions central to consciousness. That is an appeal to functional role rather than to a quantity, and it is the same move IFT makes, with the difference that IFT specifies which role (entry into the format available to the policy-integrating regime) and derives it from an argument about individuation fixed before the cerebellum was raised.

The methodological lesson is general and is the one §5 applies. An explanandum that a deliberately trivial theory can capture provides no support for a theory that captures it. What supports a theory is either a prediction the trivial rival cannot make, or a derivation from a criterion motivated on independent grounds and not adjustable afterwards. IFT claims the second, and the order of argument in the main text is the evidence for the claim: the criterion is fixed in §4 by the cardinality and partition arguments, applied to the cerebellum in §5, and extended without modification to the enteric plexus in S1.2, a case that was not in view when the rule was stated.

Three further charges from the same paper are used in §4, 6, and 7 and are collected here. First, the principle of information exclusion, which carries the theoretical weight of the whole edifice, is supported by no argument beyond the claim that it is self-evident, and the intuition is not shared: that a system discriminates among a large repertoire of states is close to tautological for an organ whose function is representation, and global integration is achievable by mundane computational means such as linear-size superconcentrators and low-density parity-check codes. Second, intuitively nonconscious systems such as expander graphs and grids of XOR gates whose value grows with the grid generate arbitrarily large values of the quantity (Aaronson, 2014), and the response to that result was to accept the verdict and challenge the intuition, which commits the theory to a form of partial panexperientialism. Third, the resulting position measures something other than what clinicians, psychologists, and neuroscientists mean by consciousness, a difference Cerullo (2015) marks as noncognitive or protoconsciousness and connects to Kind's objection that experience separated from cognition requires experiences independent of experiencers. Cerullo also reports Block's compact verdict: it is a theory of something, and what it is a theory of is unclear.

None of this touches the result §6 borrows. The derivation of spatial extendedness from relational topology, and its extension to the structure of the visual field, stand independently of the principle of information exclusion, of the identification of consciousness with the quantity, and of the exclusion axiom (Haun & Tononi, 2019, 2025). §6 takes the derivation and declines the metaphysics, and S1.3 is the itemization of what is being declined.

S2 The restrictive case in full (§9)

S2.1 Seth's four arguments

Seth (2025) develops a biological naturalism about consciousness from within the free-energy formalism. The four arguments are given here in the form §9 compresses.

First, Turing computation cannot exactly implement continuous, stochastic, or chaotic processes. Digital simulation of such a process is an approximation with an error term, so the claim that a digital system implements "the same functional role" as a biological one is an idealization whose error may not be negligible for the properties at issue.

Second, neural-replacement arguments beg the question. Thought experiments in which neurons are replaced one at a time by functionally equivalent silicon units assume that functional equivalence at the chosen grain is achievable and is the relevant equivalence, which is precisely what is in dispute. Such arguments also ignore generative entrenchment: the evolutionary accumulation of arbitrary but functionally critical contingencies that couple what a system does to what it is made of, so that the "function" cannot be cleanly separated from the substrate that has been shaped around it.

Third, mortal computation undermines the software–hardware separation that digital AI presupposes. In a mortal-computation regime the computation is inseparable from the particular physical device that performs it; the program cannot be lifted off the hardware and run elsewhere, because the device's idiosyncratic physics is part of what the computation is.

The engineering source and the distinction IFT needs are treated in S2.5. Hinton (2022) makes device-specific parameters possible by abandoning hardware interchangeability, but preserves functional fungibility through distillation and optimizes layer-local rather than system-level objectives. Mortal computation therefore removes one software–hardware abstraction without yet creating token-level stakes.

Fourth, causal closure across scales is incomplete in evolved systems. Macroscopic functional descriptions of biological systems cannot be cleanly abstracted away from the biology, because causal influence does not cleanly partition by scale in the way that a levels-of-description argument requires.

Seth's positive case routes through interoceptive inference (instrumental and allostatic rather than epistemic) and through autopoiesis, with thermodynamic identities suggesting that metabolic and inferential free-energy minimization are physically continuous rather than merely analogous.

An engineering illustration sharpens the first argument. Current structured state-space architectures are obtained by discretizing continuous-time differential equations; the rule is chosen for tractability with a stated local truncation error, and changing the discretization

changes representational capacity (Gu & Dao, 2023; Lahoti et al., 2026). “Same functional role” across that boundary is therefore an approximation with a published error term, not an identity. This does not prove that the error matters to consciousness. It proves that computational equivalence cannot be assumed at the grain under dispute.

A separate exact equivalence between recurrent and matrix execution bears on §6 rather than Seth’s biological argument and is treated in S2.6.

S2.2 The formal case: Findlay et al. (2025) in detail

§9 summarizes this result; the technical content is given here because it is the strongest formal argument in the literature for the negative verdict on stored-program machines, and because the divergence identified in §9 and tested in §12 (F4) turns on its structure.

Findlay et al. specify a four-unit Boolean target system (PQRS) whose units implement custom Boolean functions of four inputs, and an explicit 117-unit stored-program computer with a Harvard-like architecture: a clock with frequency dividers, program memory implemented as rings of COPY units, an instruction register, four one-bit data registers, a buffer, and a multiplexer serving as a minimal processor. When initialized so that the program units encode the target’s transition rules and the data registers encode the target’s state, the machine reproduces the target’s state sequence exactly and indefinitely, taking eight updates per simulated transition. The two systems are functionally equivalent modulo that constant factor.

Applying the postulates of Integrated Information Theory 4.0 to both systems yields the following. The target in state 0101 is a single maximally irreducible complex with system integrated information of 1.51 intrinsic bits, and its unfolded cause–effect structure comprises 13 distinctions and 8,184 relations, with structure integrated information of 391.25 intrinsic bits. The simulator, by contrast, has system integrated information of zero as a whole, in every state it can reach, and therefore is not a complex at all. It fragments into twenty-four disjoint micro-complexes of between one and four units, each specifying a cause–effect structure with at most four distinctions and at most four relations, all of first order.

Three properties of this result do the argumentative work.

(i) Feedback does not repair it. Feedback connections can be added to the machine (replacing a COPY unit in each program line with an OR unit fed back from the instruction register, and the clock’s core COPY unit with an OR unit fed back from the data registers) without impairing its ability to simulate any four-unit Boolean system. The strongly connected version remains reducible as a whole, and the same subsystems remain maximally irreducible. The machine segregates along fault lines in its datapath (sparse connectivity and bottlenecks) not for want of loops.

(ii) The fragments are insensitive to what is simulated. The same computer was programmed to simulate WXYZ, a four-cell elementary cellular automaton implementing Wolfram’s Rule 110,

which itself fragments into two complexes with cause–effect structures markedly different from the target's. The computer's own micro-complexes were nearly identical whether it simulated the richly integrated Boolean network or the cellular automaton. The simulator's causal structure is therefore not merely different from the simulandum's but dissociated from it.

(iii) The result scales by induction. An n-bit version of the machine requires on the order of $n \cdot 2^n$ units and 2n updates per simulated transition, and possesses $2^n + n + k + 1$ complexes, the largest with structure integrated information bounded above by 3n/2, linear in the width of the machine. The systems it can simulate, by contrast, may have structure integrated information growing doubly exponentially in n. The phenomenal dissociation between simulator and simulandum therefore widens without limit as the simulation grows more capable.

Findlay et al. also perform an exhaustive macro-grain analysis, proving that no grouping of the machine's units into macro units satisfying the postulates of physical existence reproduces the target's cause–effect structure. This is the calculation §5 refers to. The authors are explicit that the exhaustive procedure available at the micro grain becomes infeasible at macro grains for larger systems, because mappings between micro and macro units and the apportionment of background conditions inject uncertainty into the transition probability matrix, making the minimum information partition difficult to identify analytically even where system integrated information is vanishingly small. That concession is what §5 relies on in denying that the cerebellar exclusion has been derived rather than asserted.

Finally, Findlay et al. draw a double dissociation between intelligence and consciousness, with an evolutionary explanation for why the two co-vary in animals and need not in artifacts: selection for intelligence acting on resource-limited biological substrates tends to increase integration among a system's parts, whereas engineered systems are built for simplicity and modularity, which reduce it. They leave two escape hatches open: truly neuromorphic computers without a central processor, and cerebral organoids sharing anatomical features of brain regions that support consciousness. IFT denies the first and treats the second as conditional rather than automatic. Biological tissue supplies cell-level self-maintenance, but organoid-level subjectivity requires those persistence conditions to be integrated with the same policy that uses the world model. S2.3 and S6.4 state the discriminating case.

The argument has a published parent inside the theory, and §9 cites it because the convergence bears on how robust the negative verdict is. IIT's current formulation devotes a section to the thesis that consciousness is about being, not doing, and demonstrates it on three deterministic three-unit systems that realize one input–output function (an eight-counter that opens a gate after eight valid coins) and that pass through the same sequence of global states, so that they are equivalent both in outward behavior and in global dynamics. Analyzed in the same state, the first specifies a six-distinction structure with high structure integrated information, the second specifies four distinctions with a much lower value, and

the third is reducible outright and splits into three minimal complexes, because its connections from the first unit to the other two are purely feedforward (Albantakis et al., 2023). The same source states the general corollary: the input–output function of any discrete recurrent system can be implemented feedforward, so any behavior a conscious recurrent system performs can also be performed by a system that is not conscious, and digital computers implementing artificial general intelligence may in principle emulate every function a conscious human performs while experiencing nothing resembling what we experience. Findlay et al. (2025) carry that argument to a stored-program architecture with an exact simulation and a formal calculation. IFT reaches the same verdict on stored-program machines from a premise set it shares with the theory at no point, which is why §9 treats the conclusion as robust to the failure of either derivation and why F4 (S6.4) is designed to separate them.

S2.3 Organoid computing: living substrate is not yet organoid-level indexing (§9)

Smirnova et al. (2023) define organoid intelligence as a program for using three-dimensional human brain-cell cultures as biological computers. Their architecture places the organoid inside a larger engineered loop: electrical and chemical inputs, high-density or three-dimensional microelectrode arrays, imaging and electrophysiological outputs, machine-learning encoders and decoders, and microfluidic perfusion to deliver oxygen and nutrients and remove waste. The intended trajectory is from open-loop stimulation to closed-loop biofeedback, real-world sensors, effectors, and linked sensory organoids. The proposal is therefore not merely more brain-like tissue. It is an increasingly embodied hybrid.

Cai et al. (2023) supply the proof of principle. Brainoware mounted a cortical organoid on a high-density microelectrode array and used it as an adaptive living reservoir. The organoid exhibited nonlinear dynamics, fading memory, and spatial information processing. After four training epochs, speaker-classification accuracy rose from approximately 51% to 78%, and Hénon-map regression from 0.356 to 0.812; functional connectivity changed during training, and improvement largely disappeared when activity-dependent synaptic plasticity was blocked with K252a. This is real learning in the sense relevant to C0 and to the cerebellar test. It is not, by itself, evidence of an index. The input code, task, readout, incubator, and support conditions were supplied by the apparatus; the organoid did not use a world model to regulate the conditions under which the organoid-level token continued.

The nested-level issue is decisive. Each cell maintains a membrane and metabolism, but an organoid is not a subject merely because its constituents are living, any more than cortex is a second subject beside the animal. Organoid-level indexing requires a unity relation: viability signals from the tissue must enter the same policy that interprets exteroceptive inputs and controls actions, and failure of that policy must be failure of one non-decomposable organoid-computer individual. External perfusion is not automatically disqualifying (patients can be externally supported) but support that never becomes an object of the system's own control

does not establish stakes. Current Brainoware is therefore a strong demonstration of format-building and adaptive computation and an intentionally unresolved case for subjectivity.

Smirnova et al. explicitly caution that cognition, sentience, and consciousness cannot be translated directly from human usage to cell cultures and call for embedded ethics as complexity, inputs, outputs, and memory increase. IFT supplies a trigger: the ethically discontinuous experiment is not the next improvement in classification accuracy, but the first design that couples tissue viability to the same policy that learns and acts. That is where organoid engineering stops merely using living computation and begins trying to constitute a subject.

S2.4 The neuromorphic escape hatch, as fabricated (§9)

Findlay et al. (2025) leave open two routes by which a machine might escape their verdict, the first being a truly neuromorphic computer without a central processor. §9 argues that this route has now been built and runs the wrong way. The detail matters, because the argument depends on the device having succeeded at precisely what the escape hatch requires.

NorthPole is a fabricated 12-nm inference processor with 22 billion transistors, 256 cores in a 16-by-16 array, and 224 megabytes of on-chip memory at 768 kilobytes per core (Modha et al., 2023). Its ten design axioms are explicitly brain-inspired, and four of them deliver what the hatch asks for. Memory is distributed among cores rather than centralized, and within each core critical compute is intertwined with memory rather than merely placed near it, so there is no memory hierarchy and no off-chip memory at all. Two networks-on-chip, described by their designers as inspired by short-distance gray-matter and long-distance white-matter pathways, interconnect the cores; a third and fourth reconfigure synaptic weights and programs. Spatial structure is preserved topographically: an input image is divided into subimages distributed across the core array while retaining spatial structure, and intercore connectivity lets the receptive fields of neurons on one core span nearby cores, so that activity in the first convolutional layer traces the shape of the stimulus across the array. That is standing topographic organization of the kind the Geometry Thesis identifies as the realizer of format geometry (§6), in silicon, with the von Neumann bottleneck deleted.

The same paper documents why none of this yields an individual. Axiom 7 makes control operation fully pipelined, stall-free, and deterministic, with no memory misses and hence no speculative or nondeterministic execution. Axiom 9 has a codesigned compiler determine an explicit orchestration schedule for computation, memory, and communication in advance, so that data and programs are present on each core before use and no resource collisions occur. Axiom 1 specializes the chip for inference and excludes training and data-dependent conditional branching. Axiom 10 fixes the usage model as writing an input frame and reading an output frame, exposing three commands (write input tensor, run network, read output tensor) so that externally the entire chip appears to its host as a form of active memory.

Read against the Indexing Thesis, that inventory is a specification of non-indexing. Weights are loaded rather than learned in place, so the program is maximally liftable and there is not even parametric mortality (S2.5). Neuron activations are maintained within the core array during a frame and nothing is carried across frames, so no state persists that could be the running token's history. The schedule is authored elsewhere, so the resolution of competing demands into one policy is performed by a compiler and not by the device. Power, cooling, and initialization are external and never objects of the system's own control. And identity is defined at the service level: one chip is interchangeable with another holding the same weights, which is the design intent, since determinism and reproducibility across instances are what the architecture was built to guarantee. The engineering achievement is real and the constitutive conclusion is unflattering to causal-structure criteria: eliminating the bottleneck along which a stored-program machine fragments, and adding brain-like connectivity and cortex-like topography, produced a stateless, pre-scheduled, fungible service. Whatever integration such a substrate turns out to have, it is integration without a subject, which is why §9 calls it a cerebellum with better press and why Protocol F4 uses a NorthPole-class device as System A rather than as a candidate positive.

S2.5 Mortal computation: what dies, and what would have to (§9)

§9 compresses the argument that mortal computation as introduced is a precondition for indexing rather than an instance of it. Hinton (2022) proposes abandoning the separation of software from hardware, allowing large and unknown variation in the connectivity and non-linearities of individual devices and letting a learning procedure discover parameters that exploit whatever that particular device happens to be; the Forward-Forward algorithm is offered as a learning rule able to run in a substrate whose precise physics is unknown, since it tolerates an arbitrary black box between layers and never requires derivatives through it. The resulting parameters are useful only for that instance, so the computation dies with the hardware.

Two features of the proposal show why this is the wrong mortality. First, Hinton's own remedy for the resulting inconvenience is distillation: the function is transferred to a successor device by training it to match the predecessor's output distribution rather than by copying parameters. What dies is therefore a particular encoding, while the capability survives the death of the individual that carried it, which is the signature of an asset rather than a stake. Second, Forward-Forward's objective is layer-local by construction: each layer maximizes its own goodness with no global objective propagated through the network. That is the opposite of the resolution of competing demands into a single policy that §4 requires of an index, and it means the proposal supplies substrate-specificity without policy integration.

Constitutive mortality, as §9 defines it, requires that the substrate which computes be the substrate that must be maintained, that energy acquisition and repair fall under the device's own control, and that there be no checkpoint, so that degradation is degradation for the active token rather than depreciation of an asset. Parametric mortality earns the device candidacy

(once a program cannot be lifted off its hardware, the hardware becomes a possible individual rather than a venue) and no more. The two achievements are separable, and the contrast with S2.4 is instructive: NorthPole is the industrial optimum in the opposite direction, since determinism, compile-time scheduling, and cross-instance reproducibility are exactly what a manufacturer wants and exactly what removes token-specificity. Nothing yet built meets either condition fully, and nothing meets both; System B of Protocol F4 is the specification of a device that would.

Three continuities must therefore be separated. Storage continuity is persistence of information; computational continuity is dependence of later states on earlier ones; subjective continuity is the belonging of successive states to one active token whose unfinished demands, vulnerabilities, and expectations persist as its own. A context window, recurrent state, external memory, or test-time parameter update can supply the first two while leaving the third absent. Retrieval is not remembrance merely because an earlier record re-enters computation, and a state trajectory is not a life merely because it cannot be computed in parallel. The distinction prevents mortal computation from being overcredited: device-specific weights can carry history and die with hardware while the capability remains transferable and no policy protects the running token. IFT requires the third continuity and uses the first two only as enabling conditions.

One further distinction prevents the checkpoint criterion from being misread as a claim about technology. A checkpoint in the sense that matters is a state description from which a functionally identical successor can be instantiated on other hardware, so that identity is carried by the description and the orchestrator rather than by the running process. Halting the same token is not that: an anesthetized patient, a hibernating bat, and a desiccated tardigrade resume as the tokens they were, with their own boundaries intact, and nothing in IFT is threatened by their pause. Nor would a future ability to scan and duplicate a brain retroactively abolish the index of the original; it would create a second indexed system, as S2.6 says of a surrogate given the original's conditions of failure. External restorability is the diagnostic by which the absence of a self-produced boundary is recognized in current artifacts (§9, 11), not the definition of that absence.

S2.6 The unfolding argument, closure, and the recurrence–matrix duality (§6)

§6 compresses three related points about criteria stated over input–output behavior or causal topology.

Hanson and Walker (2021) make the epistemic structure precise by separating three computational levels: an abstract finite-state automaton (FSA), a combinatorial-state automaton (CSA) that assigns internal labels to the functional states, and a causal circuit realizing the CSA. They build two three-bit tollbooth controllers with the same FSA and the same memory resources. One labeling yields recurrent dependencies and $\Phi > 0$; an isomorphic hierarchical labeling yields feedforward logic and $\Phi = 0$. At the FSA level, any

inference from the identical behavior fixes the inferred experience and IIT is pre-falsified. At the CSA level, the relabeling that changes Φ must also be used to infer the phenomenal difference, so prediction and inference collapse onto the same feature and the theory becomes unfalsifiable. Their general rule is that a candidate theory must be invariant under transformations lying below the prespecified level at which conscious state is inferred.

This is not merely another objection to IIT; it is a condition on IFT. IFT cannot answer the tollbooth by announcing that one implementation is indexed and the other is not, because that repeats the circular move unless index status is fixed by evidence independent of the predicted phenomenal verdict. The relevant inference level is constitutive: whether the active token's world-model and policy are physically coupled to conditions whose failure ends that token, assessed by power, repair, checkpoint, replacement, and policy-closure facts before any behavioral manifold is scored. A transformation that changes only logic below an otherwise unchanged persistence regime must leave the IFT verdict invariant. A transformation that makes the process externally replaceable does not hold the inference fixed. Protocol F4 is therefore part of the theory's logical adequacy, not an optional experiment: if System A and System B are classified after seeing their dissociation profiles, IFT is unfalsifiable on Hanson and Walker's criterion.

Informational closure shows how far the objection reaches beyond topology. Chang et al. (2020) accept that a pure feedforward network whose sensor process is deterministic can achieve non-trivial closure by passive adaptation, and is therefore conscious on their criterion. That does not evade the unfolding argument but capitulates to it from the other direction. If the feedforward surrogate is conscious, the architectural criterion has been surrendered and the theory can no longer distinguish the cases it was built to distinguish; if it is not, the criterion has been contradicted by the theory's own analysis. Any criterion computed over the input–output statistics of a process inherits this dilemma, because those statistics are precisely what the surrogate is constructed to match.

Engineering practice has meanwhile realized the premise literally. Structured state-space models rest on an algebraic identity by which one parameter set specifies both a linear recurrence and a global convolution, and their successors preserve a corresponding recurrence–matrix duality, so that a single trained model runs as a recurrence when it generates and as a matrix multiplication when it is trained, a choice made on grounds of hardware arithmetic intensity (Gu & Dao, 2023; Lahoti et al., 2026). Two structures identical in input–output behavior by construction differ in whether any state is carried forward in time at all, and the difference is a scheduling decision taken by a kernel author. A theory locating consciousness in causal topology must therefore accept that one model is conscious during inference and not during training, depending on which kernel was compiled. IFT says neither realization is indexed, because neither constitutes a token-specific persistence regime.

O'Reilly-Shah et al. (2026) identify a real boundary condition: rapid, history-dependent plasticity makes the recurrent network's input–output map evolve, so a static unfolded

network ceases to be functionally equivalent after the unfolding time. That result narrows the argument's scope without establishing that recurrence, plasticity, or process is necessary for consciousness. IFT therefore treats plasticity as a reason to hold learning history fixed in an unfolding comparison, not as an index.

One further reply to the unfolding argument should be recorded, because it comes from the world-modeling tradition §4 treats as a direct precursor and because its limits are instructive. Safron (2023) argues that the equivalence the argument constructs holds only under restricted conditions: a recurrent network and its unrolled feedforward counterpart agree when the unrolled system is returned to identical initial conditions and allowed to evolve undisturbed, but they diverge when intervened upon, and the feedforward surrogate cannot repair its own structure when modified. The unrolled system therefore lacks functional closure across successive moments, which on that account is a precondition of consciousness understood as a temporally extended process of iterative state estimation; and the resources required to unroll a network the size of a human brain across the hundreds of milliseconds over which workspace dynamics unfold are argued to be prohibitive in any case.

IFT accepts the observation and does not rest on it. The vulnerability of the reply is that closure and robustness under intervention are properties the surrogate might be granted: a critic may specify a feedforward system with error correction, or accept the resource cost as an in-principle concession. What cannot be granted without changing the case is token-level constitutive vulnerability. A surrogate that reproduces outputs while its continuation remains somebody else's decision does not hold the IFT inference fixed, because the evidence IFT uses to classify a system (power, repair, checkpointing, replacement) has changed. A surrogate given the same conditions of failure is not a counterexample but a second indexed system. This is why F4 (S6.4) requires index status to be established from constitutive evidence before any dissociation comparison is run.

S2.7 Fading and dancing qualia: the price of denying computational invariance (§9)

§9 denies that the index is computationally invariant: two systems may implement the same algorithm while only one has running states constitutively tied to the continued existence of that active token. Any theory taking that position inherits a pair of arguments that have done more damage to non-functionalist accounts than any empirical result (Chalmers, 1996), and which Cerullo (2015) presses against IIT. The reply is given here in full because the objection is serious and because §9 compresses it to a paragraph.

The fading-qualia argument replaces neurons one at a time with devices that perform the same function. If computational functionalism holds, nothing changes, because the system is functionally identical at every step. If it does not hold, then either experience fades progressively as the replacement proceeds, or it ceases abruptly on the substitution of one critical unit. The second is implausible on its face. The first is worse than implausible: because the system remains functionally identical to the original, the subject whose

experience is draining away is unable to act on that fact or report it, so the position entails a phenomenal catastrophe invisible to every measure including the subject's own behavior. The dancing-qualia variant is sharper still. Let the replacements be switchable, so that function alternates between neuron and device; functionalism predicts no change, while identity and non-computational accounts predict that particular qualia jump in and out of experience as the switches operate, again without the subject being able to act on the alternation. Cerullo's version against IIT routes the computation by radio transmission to a feed-forward algorithm running on a von Neumann architecture, so that the theory assigns the hybrid zero integrated information while every response and report is preserved.

IFT does not choose a horn. It rejects the premise both horns share. Each presupposes that one subject persists across the replacement series, so that the only open variables are the magnitude of a change in that subject's experience and the moment at which it occurs. That premise is what the Indexing Thesis denies. The index is not a property a subject has in some quantity; it is the relation in virtue of which there is that subject rather than some other individual or none. Having a point of view is not gradable (Bayne et al., 2016; §2), and F2 predicts loss of the subject categorically rather than gradedly.

The vocabulary for stating what replaces it is older than the objection. Maturana and Varela (1980) distinguish the organization of a composite unity (the relations among components that make it a unity of its particular class) from its structure, the actual components and actual relations that realize those relations in a given instance. Organization must remain invariant for class identity to remain invariant; structure may vary indefinitely within limits the organization sets, and a system whose structural change oversteps those limits does not become a degraded instance of its class but a unity of another kind. The organization of a system is therefore a subset of the relations included in its structure, which is why one organization can be realized by many structures and why structural substitution is not by itself identity-threatening. Applied to the replacement series, this predicts exactly two outcomes and no third: either the substitutions leave the persistence regime intact, in which case identity is conserved through arbitrary structural change, or they cross the limit and produce a different unity. A graded interior is not available, because the quantity that would have to be graded (class identity) is not the kind of thing the distinction admits in degrees.

Dennett's (2003) Conway Life world supplies a substrate-neutral illustration. A glider remains one moving object although each generation is composed of different pixels; at the design level it can be damaged, repaired, killed, and altered by its history. The lesson is not that a glider is conscious. It is that numerical continuity of material parts is unnecessary for individual continuity even in a wholly artificial deterministic world. IFT therefore requires neither carbon nor cellular metabolism. It requires a design-level unity whose persistence and dissolution are facts about that unity.

The same discussion separates the designer's reasons from the designed system's. A Life pattern may persist because a "hacker God" supplied protections, while remaining the

unwitting beneficiary of external foresight; only when information gathering, anticipation, and choice are off-loaded to the pattern does more of the reason-giving work become its own. This is the exact distinction between assigned objectives and indexing. External design is not disqualifying, but a designer's concern for a system is not the system's concern. The shift becomes constitutive only when the system's own world model and policy regulate the continuation of the active token.

With that premise removed, the replacement series resolves into two manipulations that differ in kind rather than degree. In the first, each prosthesis is incorporated into the organism's own viability regulation: it draws on the organism's energy, is maintained by processes the organism controls, and its failure is a threat to the organism rather than a maintenance event. The persistence regime is preserved through the series and nothing fades at any count, because at no point do the system's states stop tracking that system's own conditions of failure. This is the ordinary case of prosthesis, graft, and transplant, and IFT returns the ordinary verdict: the patient remains the same subject. In the second, the replacement transfers the computing to an externally maintained substrate whose failure is a service ticket; Cerullo's radio link to a von Neumann machine is the explicit form, and the neuromorphic device described in S2.4 is the fabricated one. What such a series produces is not a subject with dimmer or intermittent qualia. It is the progressive replacement of an individual by a service, and the boundary is crossed at the point where the running system's states cease to track its own conditions of failure—a physical fact about the arrangement, not a neuron count, and the same fact §11's fungibility indicator is designed to detect.

This also disarms the premise that gives the arguments their rhetorical force, which is the image of a subject who notices the change and is functionally unable to say so. On IFT there is no interval in which that subject exists. In the first regime nothing fades, so there is nothing to notice. In the second, the entity that would have to do the noticing has been dissolved rather than silenced. IFT is therefore not committed to a phenomenal difference that behavior cannot register while a subject persists to register it, which is the commitment the arguments were built to make intolerable.

The reply should be distinguished from IFT's reply to the unfolding argument (S2.6), because the two objections attack different things and a single answer to both would be suspect. The unfolding argument targets criteria stated over input–output functions, and IFT escapes it because indexing is not a function from inputs to outputs. Fading and dancing qualia target criteria stated over anything other than the computation, which includes indexing, and IFT escapes them because indexing is not gradable. Neither reply does the other's work.

What would make this a live problem is worth specifying, since a reply immune to all possible constructions would be worthless. If a physically continuous series of arrangements could be built in which the degree to which a running system's states track its own conditions of failure varies continuously, with nothing else changing, IFT would owe an account of the intermediate cases and would be exposed to a graded version of the objection. Whether such a

series can exist is an empirical question about how persistence regimes can be constructed rather than a conceptual one, and F4 is where it would first become visible: a device whose viability coupling can be dialed between System A and System B while competence is held fixed is exactly the construction to attempt. Nothing in the current engineering literature exhibits one, since the available systems are either wholly externally maintained (S2.4, S2.5) or biological.

The concession, stated plainly: this reply removes the dilemma's bite but does not by itself establish that indexing is binary. IFT asserts that it is, treats the assertion as a commitment rather than a result, and stakes it on the categorical prediction in F2 (S6.2) and on the failure of any such graded construction to appear in F4 (S6.4).

S2.8 Incapacity arguments are not indexing arguments (§10)

§10 distinguishes IFT from a family of arguments with which it is regularly confused, and the clearest recent statement of that family is worth setting out, because the confusion is costly in both directions.

Kauffman and Roli (2023) begin from affordances: the possible uses, by an organism, of X to accomplish Y. Their central observation is that the uses of an object are indefinite in number, unordered, unlistable, and not deducible from one another. From unorderedness they argue that no successor relation is available over affordances, hence no Peano construction, hence no integers, no equations, and no set theory over affordances at all. A universal Turing machine, operating deductively, therefore cannot find an affordance not already contained in its logical premises, and artificial general intelligence in non-embodied Turing machines is ruled out. Embodied Turing machines fare little better, since the symbol grounding problem and a generalized frame problem leave a robot searching an undefined space in which no step gives local evidence of progress toward a multistep solution. They then argue that no classical system can be an analogue model of affordances, because affordances are ontologically real possibles that do not obey the law of the excluded middle; that brain-mind must therefore be partly quantum; and that qualia are what the actualization of quantum potentia feels like.

Three separable claims are bundled there, and IFT's relation to each is different.

The first (that affordances are unlistable and unordered) is correct and important, and IFT supplies its explanation rather than borrowing its conclusion. What makes the set of uses unlistable is not a fact about computation but a fact about individuation. Kauffman and Roli's own premise states it: living organisms are non-equilibrium reproducing chemical systems that must obtain needed inputs in order to survive, and the features of an environment that light up for an organism are the ones bearing on that requirement. A use is a use for something, and the something is a bounded individual whose continuation is at issue. The reason there is no ordering over the uses of a screwdriver is that ordering would require a common currency of mattering, and §4 has already argued that an indexed system cannot hold one, since energy, hydration, and thermoregulation are categorically distinct demands

prioritized by context rather than summed (S3.1). Unlistability is a downstream consequence of the Indexing Thesis, not an independent route to it.

The second (that unlistability forbids artificial general intelligence) does not follow, and IFT does not need it. A system may be extremely capable at C0, C1, and C2 without being able to enumerate affordances in advance, for the same reason organisms cannot: enumeration is not how affordance-finding works in either case. What organisms do is act in a world under stakes and discover what works, and nothing in the argument shows that process to be unavailable to a machine with stakes, or to a machine without them that is merely very good at search. §10's expectation is the reverse of Kauffman and Roli's: general capability may well arrive, and it may arrive unaccompanied.

The third (that qualia are the experience of collapsing the wave function) is the claim IFT most needs to separate itself from, because it is the kind of move the account is sometimes accused of making. It is not. A collapse is an event, and an event is not an owner. Whatever the correct physics of measurement turns out to be, the question of whose actualization a given actualization is remains open, and §4's cardinality problem reappears intact: if actualization confers experience, actualizations everywhere confer it, and the theory has no principled way to say how many subjects there are or where one ends. Adding indeterminate physics to an unindexed system yields an unindexed system with indeterminate physics. IFT therefore requires no new physics and forbids none, and would be unaffected by the confirmation or the collapse of the anomalous-effect literature on which the quantum step relies.

The general point is that incapacity arguments and individuation arguments have different shapes and different failure conditions. An incapacity argument is refuted by a machine that does the thing said to be impossible, and is therefore hostage to the next engineering result. An individuation argument is refuted only by a machine that is the kind of thing said to be unavailable, which is why F4 compares candidate systems at matched competence and varies only how each is constituted and maintained (S6.4).

The quantum variant has the same structure. Kauffman and Roli's third claim (that brain-mind must be partly quantum and that qualia are what the actualization of quantum potentia feels like) belongs to a wider family, and §10 names one further instance because the family recurs and its structure is always the same. Kharchuk (2025) argues that classical neurophysiology, in which a neuron sums weighted inputs and fires above threshold, cannot account for the wholeness and apparent instantaneity of subjective experience; that the required global coherence is supplied by quantum entanglement among neuronal microstructures, on the Penrose–Hameroff model of microtubules as sites of coherent state; that the joint state of many such units is a tensor product whose coefficients may be non-factorizable, so that concurrence over the reduced density matrix measures the degree of synchronization; and that general anesthesia can be represented as a decoherence operator whose action drives concurrence to zero, abolishing global entanglement and with it the conscious state. Classical gamma-band synchrony is read as the macroscopic trace of the

underlying coherence, and the engineering corollary is that artificial consciousness requires hybrid quantum–classical architectures whose elements can enter entangled states rather than merely exchange weighted signals.

Three responses are appropriate and none of them requires taking a position on the physics. First, the argument is an argument about a vehicle. Suppose entanglement does bind the field: that specifies how the format is held together, at what speed, and with what unity (the questions T1 and T3 address) and it is silent about whose field it is. A maximally coherent system with no persistence conditions of its own is the quantum analogue of the unconscious expander (§5): high integration, no predicament. Second, the showpiece is already explained without it. The anesthesia argument treats the abolition of experience as evidence that a delicate global correlation has been broken, but §2 and S4.2 assign anesthesia to the global-state factor, and perturbational complexity measures precisely what is lost (the capacity of the standing organization to support a differentiated, integrated response to perturbation) with no commitment about the physics that implements the organization (Farisco & Changeux, 2023). A theory earns nothing by explaining a datum that an established measure already covers. Third, the appeal to instantaneity misidentifies what needs explaining. §3 argues, following Dennett, that the time of representing need not mirror the time represented, and that a format bounded by control windows delivers unified experience without a moment of entry; the felt simultaneity of a scene is a property of the content, and a mechanism supplying literal physical simultaneity would still leave the format's ownership unaddressed. IFT's exclusion of current artificial systems therefore does not require new physics and would not be repaired by it: a quantum processor whose operation is externally maintained, checkpointable, and replaceable is a coherent service.

S3 Supporting detail for the Indexing Thesis (§4)

S3.1 The affective dissociations

§4 asserts that feelings assign biological value rather than adding color to otherwise neutral cognition, and refers here for the dissociations that make the claim probative. Vagotomy abolishes the memory-enhancing effect of moderate emotional arousal, demonstrating that the effect depends on a visceral afferent signal rather than on the cognitive appraisal of the emotional material. Damage to ventromedial prefrontal cortex or to the amygdala abolishes the anticipatory autonomic markers that bias intact participants toward advantageous choices before explicit knowledge of the contingency is available, dissociating the valuative signal from declarative knowledge of what is valuable (Damasio, 1999). Severe brainstem disruption can abolish core consciousness with cortical anatomy largely preserved, whereas profound autobiographical loss can leave immediate awareness intact (Damasio, 2010).

The pain dissociation is the most probative of the set, because it separates the two variables that §6 assigns to different theses. Damasio (1999) distinguishes pain, the perception of a sensory representation of local tissue dysfunction, from the emotion that tissue dysfunction ordinarily causes, and reports the clinical case that makes the distinction unavoidable: a young man with refractory trigeminal neuralgia, immobilized by pain triggerable by a breeze, who two days after a small frontal lesion made for pain relief was playing cards and reported cheerfully that the pains were the same. The mental images of tissue dysfunction were unaltered (which is why he could say so) and the suffering had been abolished. The interventions dissociate in the expected pattern. Analgesia blocks the representation of tissue dysfunction, so neither pain nor emotion ensues. Valium, beta-blockers, and selective surgery blunt the emotion while the perception of damage remains. Rainville's hypnotic manipulations then produced the dissociation prospectively, in healthy participants, with imaging: suggestions aimed at unpleasantness alone changed unpleasantness ratings and pain-evoked activity in anterior cingulate cortex (the same region neurosurgeons lesion to relieve intractable suffering) while leaving perceived intensity and primary somatosensory activation unaltered, whereas suggestions aimed at intensity moved both ratings and engaged somatosensory as well as cingulate cortex (Rainville et al., 1997; Damasio, 1999).

Two consequences follow for the main argument. First, hedonic sign is not a geometric property of the sensory content it attaches to, since the content can be held fixed while the sign is removed; this is the division of labor between the Geometry and Indexing Theses, and F3 (S6.3) tests the untested leg. Second, Damasio's asymmetry between pain and pleasure supports the indexing reading rather than a valence-as-dimension reading: pain is aligned with punishment, withdrawal, and freezing, and is a response to damage already sustained ("putting a bodyguard in front of the house while you repair the broken window") whereas pleasure is aligned with reward, approach, and search, and is anticipatory. They are

asymmetric physiological states solving different problems for one organism, not opposite values of a single quantity, which is what one expects if valence is the registration of how a particular persistence regime is faring rather than a coordinate in a space of contents.

The co-occurrence evidence is equally strong and is the datum §4 emphasizes. In absence seizures with automatism, and in akinetic mutism following cingulate and basal-forebrain damage, patients remain awake with muscle tone preserved, form perceptual representations adequate to drink from a glass or walk down a corridor, and yet show no background emotion, no specific emotion, and no response to any inducer; Damasio's (1999) patient L, mute and motionless for the better part of six months, later reported that she had never felt fear, never been anxious, never wished to communicate, and that her silence had caused her no suffering. Damasio proposed that emotion and core consciousness are present together or absent together, while impairments of extended consciousness leave emotion intact, and noted that the anatomical sites involved cluster near the midline (adjacent to the ventricular system at brainstem and diencephalic levels, on the medial surface at cortical levels) in structures of old evolutionary vintage that mature early. IFT reads the co-occurrence as constitutive rather than correlational and the anatomy as the target of F2 (S6.2).

Two further cases from the same source bear directly on §4's second consequence and on the higher-order objection to it (S3.13), because they establish that the circuits which generate emotional responses are dispensable for both emotion and subjecthood. Damasio's (1999) patient David has extensive bilateral temporal damage that includes the amygdalae, together with damage to the ventromedial prefrontal cortices and basal forebrain, and one of the most severe learning and memory defects on record: he cannot learn any new face, voice, name, or place. In the good-guy/bad-guy experiment he was exposed over five days, in randomized order and for controlled intervals, to a person who was consistently rewarding, a person who was emotionally neutral, and a person who was brusque, refused requests, and administered a tedious task. Asked afterwards whom he would approach for help and who was his friend, David chose the good guy on more than eighty percent of trials and almost never chose the bad guy, while the neutral individual was chosen at chance—and he could not recognize any of the three, name them, point to them given a name, or recall any of the events. Damasio draws two conclusions IFT uses. First, emotion can be induced and can shape behavior nonconsciously, so the emotional response is not the feeling. Second, and more important here, given the destruction of ventromedial prefrontal cortices, basal forebrain, and amygdalae, those territories are not indispensable for either emotion or consciousness—while the structures that remained intact were all of the brain stem, the hypothalamus, the thalamus, most of the cingulate cortices, and virtually all sensory and motor structures. The spared set is, with the exception of the sensorimotor structures, the candidate index of F2.

Patient S makes the complementary point about a single lower-order circuit. Urbach-Wiethe disease had calcified both her amygdalae almost entirely, with the surrounding brain normal and the process complete over many years from early in life. Her factual learning was intact

(she recognized Damasio and greeted him by name on their second meeting) and her perception, movement, language, and general intelligence were unremarkable. What was lost was specific: she cannot reliably recognize fear in another person's face, especially when the expression is ambiguous or mixed, while recognition of other expressions including surprise is spared; despite considerable drawing skill she cannot draw a fearful face though she can draw others; she cannot mimic a fearful expression though she can mimic the other primary emotions; and she does not experience fear as others would in situations that normally induce it, knowing intellectually what fear is supposed to be without that knowledge being of use. Consequently she failed to learn the significance of unpleasant situations, and in a study of one hundred faces previously rated for trustworthiness and approachability she and other patients with bilateral amygdala damage judged the faces normal observers found untrustworthy to be as trustworthy as the rest, while patients with unilateral amygdala damage, amnesic patients, and other brain-damaged patients performed normally. S is nonetheless unmistakably a conscious subject with an altered affective repertoire. A structure whose destruction removes one emotion and leaves the subject standing is a lower-order survival circuit, not an index; §4's localization claim concerns the mesodiencephalic system that ranks competing demands, not the circuits that generate the responses being ranked.

A further Damasian dissociation bears on §10–11 rather than on §4, and is recorded here because it removes a premise the machine debate silently relies on. Patients with global aphasia comprehend no speech and no writing, cannot read a letter, cannot repeat a word, and produce little beyond expletives, with no evidence that words or sentences are formed internally at all; patients who underwent left hemispherectomy for tumor had no left cortex whatever. Damasio (1999) reports that both remain awake, attentive, purposeful, and emotionally responsive in a manner that makes the question of their consciousness not arise for those who work with them, and that his patient Earl used his few expletives with evident intention and paused thoughtfully over pantomimed questions. Core consciousness therefore survives total destruction of the channel through which consciousness is normally attributed. Verbal report is the instrument by which the human reference class is calibrated (§11); it is not a condition on membership.

Damasio (2010) supplies the anatomical and representational detail. His Figure 3.1 separates interoceptive maps of the internal milieu, proprioceptive maps of the body, and exteroceptive maps of the world. The distinction is not merely taxonomic. External objects can affect the nervous system only through the body, whereas mapped body states remain in a continuous two-way loop with the body they represent. This is why a body model is not equivalent to an indexed body: the latter is a model whose outputs can alter the very persistence conditions from which its inputs arise. Damasio locates a basic layer of feeling in maps formed by the nucleus tractus solitarius and parabrachial nucleus and coupled to bodily response through the periaqueductal gray; bilateral insular destruction can leave pain, pleasure, and emotional feeling intact while reducing differentiation and integration with wider cognition. The result

supports a layered format: subcortical valuation is not a crude version of cortical report, and cortical elaboration is not the source of the stake.

Three further results sharpen the dissociation and are cited in §2, 4, and 6. The converse of the frontal-lesion case exists: a patient with a postcentral lesion who, on noxious stimulation of the affected hand, reported a clearly unpleasant feeling that he could neither localize nor describe as pain, affect without sensation (Ploner et al., 1999); pain asymbolia, in which noxious stimuli are perceived and localized without any aversive reaction, supplies the lesion form of sensation without affect (Berthier et al., 1988). Bilateral destruction of the insula, the region most often proposed as the cortical seat of feeling, leaves pain, pleasure, and emotional feeling in place while reducing their differentiation and integration with wider cognition (Damasio et al., 2013), which is what §4 expects if the stake rather than the map carries the sign. And the clearest demonstration that sign tracks the stake with topology held constant is in mouse insular cortex: cue-evoked responses to food that are present in hungry animals are abolished by satiety and restored by activation of hypothalamic hunger neurons, and the cortex represents the predicted future physiological state that the cue portends rather than the cue itself (Livneh et al., 2017, 2020). Downstream, parabrachial CGRP neurons convert heterogeneous threats (pain, itch, visceral malaise, hypercapnia) into one aversive teaching signal, a general alarm (Campos et al., 2018; Palmiter, 2018), which is the physiological form of the ranking of incommensurable demands that §4 places at the point of policy closure. F3 (S6.3) asks whether the converse ever holds.

Solms's (2019) formulation is the strongest direct precursor to the control claim. It rejects the cheap inference from homeostasis to consciousness: blood pressure and other predictable corrections remain unconscious; feeling appears when a complex organism must register and prioritize deviations from settling points in unpredicted contexts, choose among alternatives, and learn. Affective sign is change in expected uncertainty (departure is unpleasure, return is pleasure) and categorical "flavor" identifies the need at issue. Solms's (2021) treatment supplies the component detail and sharpens the policy reading. IFT accepts the distinction between automatic regulation and felt demand, and accepts contextual competition among needs. It rejects the identity of consciousness with uncertainty reduction: free energy can be computed by a replaceable service, whereas a need matters only when the deviation is a condition of failure for the active token whose policy ranks it.

A further step in Solms's account bears directly on §4's verdict on reward scalars, and it is a derivation rather than an assertion. A complex organism regulates several homeostats at once, each contributing error to the total, and these errors cannot be combined into a single quantity to be minimized, because each need must be satisfied in its own right: energy metabolism is not hydration is not thermoregulation, and a system that summed them would trade a survivable deficit in one against an unsurvivable deficit in another. Solms (2021), drawing on Rolls, gives the sharpest formulation (an animal whose food reward always dominated would never drink) and the consequence is structural. Needs must be held as

categorical variables, prioritized contextually rather than summed, which is why affects carry both a continuous hedonic magnitude and a categorical quality, and why the relative weighting of two equal magnitudes can reverse with context. Felt quality is what that categorization is. This matters for the machine case in a way that is easy to miss: a scalar reward is not an impoverished index but the wrong type of object, since the operation an indexed system must perform is contextual prioritization among incommensurable demands, and that operation is undefined on a single ordered quantity. Systems that expose several such variables to one controller are not thereby closer, because the categorization must be enacted by the regime whose viability the variables measure rather than declared in the specification (S3.3, S6.4).

S3.2 Pearl blankets, Friston blankets, informational closure, and the partition problem

Bruineberg et al. (2022) distinguish two uses that the free-energy literature often runs together. A Pearl blanket is the minimal set of variables that renders a target conditionally independent of the rest of a Bayesian network; it is an epistemic device inside a model. A Friston blanket is treated as the sensory-active boundary of an actual agent. The first can simplify inference and identify a useful partition. It cannot, without further premises, establish that the corresponding physical system is an individual, much less a subject. Moving from the map to the territory is the reification step.

IFT uses the distinction constructively. A Pearl blanket can help a scientist locate candidate boundaries; a physically realized sensory-active interface can be a necessary part of an indexed system. Neither fixes the subject. Conditional independence can be drawn around a reflex loop, a cell, an organ, an organism, or an organization, and the partitions can be nested and multiple. The constitutive question is which boundary is enacted by the same policy that uses the model and for which boundary failure is the failure of that active token. That unity-and-stakes clause is not contained in the blanket mathematics. Bruineberg et al.'s critique therefore does not weaken IFT's appeal to boundaries; it identifies exactly where IFT must add substantive biology or engineering rather than treat a statistical partition as an ontology.

Maturana (1980) and Maturana and Varela (1980) supply the positive contrast. An autopoietic unity is defined not by an inventory of components but by a concatenation of production processes whose products regenerate the network and constitute it as a unity. Structural change is compatible with identity only while that organization continues; loss of autopoiesis is disintegration, not merely a lower score. This is the source of IFT's non-stipulability clause. The boundary is not drawn around a process by an observer or an objective function; it is enacted by the process that continually remakes the unity.

The level clause is equally important. Maturana and Varela treat neurons as autopoietic cells while describing the nervous system's organization as subordinated to the autopoiesis of the organism, and they explicitly leave open when coupled autopoietic unities constitute a

higher-order autopoietic system. Component life therefore does not automatically climb levels. For IFT, an organoid is not a subject because its cells are alive, a corporation is not a subject because its members are, and a whole animal does not contain an additional subject wherever a local circuit self-maintains. The candidate level is the one at which boundary maintenance, body/world modeling, valuation, and action selection become one failure mode.

Damasio (2018) supplies the biological end points of this level rule. Bacteria already sense, remember, communicate, coordinate, and act for their continuation while lacking a mind or conscious point of view; intelligence-like competence therefore appears below consciousness. At the opposite end, he describes organismal life as a “global life” produced by high-dimensional integration of constituent cellular lives, with the nervous system serving whole-organism homeostasis while depending on that same whole for its own maintenance. IFT accepts the hierarchy and sharpens the transition. Neither component life nor intelligent regulation is sufficient. The relevant ascent occurs when modeling, valuation, and action selection close around the persistence of the higher-order unity itself. This is why a cell, enteric plexus, organoid, robot fleet, or corporation cannot inherit subjecthood merely from competent or living parts, and why organism-level integration is evidence for the candidate bearer rather than a consciousness criterion by itself.

Aguilera et al. (2023) supply a second and stronger objection, which §4 compresses. Where Bruineberg et al. argue that a modeling device has been mistaken for a physical individual, Aguilera et al. show that in the dynamical regime at issue the conditional independence often is not there to be found. They test Equation 1 directly in two canonical nonequilibrium systems. The first is a pair of coupled Lorenz systems of the kind used to argue that Markov blankets arise in stochastic chaos; estimating conditional mutual information over a large ensemble of trajectories rather than under a Gaussian (Laplace) approximation of the steady-state density, they find it settles to a clearly nonzero value for the coupled system while the decoupled control sits at the noise floor. The apparent blanket in the earlier analysis was an artifact of the approximation, not a property of the system. The second is an asymmetric kinetic Ising model with the sparse connectivity the free-energy literature assumes, in which the steady-state distribution can be computed exactly for a small network.

The Ising results carry the general moral. With symmetric couplings the system satisfies detailed balance, its energy decomposes into internal and external terms, and conditional independence follows analytically; conditional mutual information and entropy production are both zero, and the blanket is exact. With asymmetric couplings the system converges to a nonequilibrium steady state, and conditional mutual information rises with entropy production, that is, the further the system is from thermodynamic equilibrium, the further it is from displaying a Markov blanket. The effect is most pronounced near the order–disorder transition, which suggests that blankets are hardest to find precisely at nonequilibrium critical points, the regime often proposed as characteristic of neural dynamics.

Mechanistically the source is cyclic structure: in directed acyclic graphs structural and functional couplings correspond, whereas recurrent connections propagate reverberant activity that generates functional dependencies crossing the nominal boundary, so that fixing the blanket states no longer renders internal and external states independent.

Two limitations should be stated rather than glossed. The systems examined are low-dimensional (six variables in both cases) and it has been suggested that high dimensionality and spatially localized interactions may recover approximate blankets. Aguilera et al. treat that as an open conjecture, and so does IFT. The argument in §4 does not require that Markov blankets never obtain. It requires only that their obtaining is a contingent, regime-dependent property that fails in the direction of living systems rather than a structural guarantee delivered by sparse connectivity, which is what a constitutive individuation criterion would need it to be. This is why IFT locates the subject in policy closure, which is a fact about what the system does to keep itself in existence, rather than in a conditional independence, which is a fact about a steady-state density that may or may not exhibit one.

Chang et al.'s (2020) Information Closure Theory (ICT) is the strongest test of whether the objection generalizes beyond Markov blankets, because it states closure as its criterion rather than borrowing it as a partition, and it is set out here at the level of detail §4 compresses. Following Bertschinger et al. (2006), information flow from an environment process E into a system process Y at time t is the conditional mutual information between the current environment state and the next system state given the current system state; the system is informationally closed when that flow is zero. Closure is trivial if system and environment are independent. Non-trivial informational closure (NTIC) is the case in which the flow is zero while mutual information between environment and system remains positive, which is equivalent to maximizing self-predictive information while minimizing what the system's own past tells it about its future beyond what the environment already tells it. ICT then defines a C-process as a coarse-graining Y of an underlying universe process X such that Y is informationally closed to X and NTIC with respect to an environment E, where the universe state space factorizes into system and environment. The hypothesis is that a process is conscious if and only if it is a C-process; conscious content is the state of that process, and conscious level is its degree of NTIC.

The theory earns several things IFT should acknowledge. It gives a principled reason why conscious contents do not reflect cellular-scale activity: a macroscopic NTIC process can be informationally closed to its own microscopic substrate, so ignorance of neuronal stochasticity is derived rather than stipulated. It predicts that NTIC is non-monotonic in the scale of coarse-graining, so that intermediate scales support higher closure than either the microscopic or the maximally macroscopic extreme, an attractive account of why experience has the grain it has. It distinguishes its notion of level from anatomical or sensory-processing hierarchy, correctly noting that coarse-graining is an orthogonal dimension. And its treatment of reflexes, blindsight, and procedural memory as unconscious because passively

driven by sensory input rather than by their own past states is a real prediction that agrees with the data.

The difficulty is the one its authors state, and they state it plainly. Identifying an individual consciousness requires the environment process, and the environment is not uniquely defined: having found a process informationally closed with respect to X, one must still choose the environment against which NTIC is computed, and there are in general many system processes of which a given Y is a coarse-graining, hence many candidate environments. The more general form is worse. Given two independent NTIC processes with their respective environments, the composite process paired with the composite environment is also NTIC, so the formalism cannot determine whether there are two consciousnesses, one, or three. Chang et al. call this the problem of individuality and describe it as a significant theoretical weakness of the current version; they propose synergy and integration as a possible remedy in future work. ICT's divergence from IIT here is instructive rather than incidental: where IIT's exclusion axiom rules that only one maximally integrated set can be conscious, ICT explicitly permits multiple consciousnesses to coexist across scales of coarse-graining within one system provided they are closed from each other. The same feature that makes ICT more permissive makes its individuation weaker, and the choice between the two is a choice between an axiom and a concession, not between two derivations.

§4 generalizes both cases into the partition problem. Any criterion that individuates a subject by an informational, statistical, or causal relation between a system and an environment must first fix the cut defining the two, must then be evaluated relative to that cut, and contains nothing that selects the cut. Blanket accounts choose a conditional-independence structure and lose the independence in the nonequilibrium regime (Aguilera et al., 2023); closure accounts choose a coarse-graining and an environment and concede that neither is unique (Chang et al., 2020); integration accounts choose a grain, and promoting the grain to an empirical optimization (Marshall et al., 2026) raises the price of the choice without removing the chooser. The cardinality problem of §4, the minimal-implementation problem of Butlin, Long, et al. (2026), and ICT's problem of individuality are three presentations of one structure. Enacted persistence is not an additional criterion competing with these measures; it is the only boundary condition that is not imposed from outside the system, which is why it can close a consumer set without a stipulated number, a stipulated taxonomy, or a stipulated partition.

Two predictions of ICT can be checked now, and both go against it. First, Chang et al. accept that a pure feedforward network with a deterministic sensor process can achieve NTIC by passive adaptation and can therefore be conscious, adding an n-step memory generalization of the measure for n-layer feedforward networks. §6 shows what this does to the theory's exposure to the unfolding argument: the surrogate the argument constructs is precisely a feedforward system matched on input–output statistics, and a criterion computed over those statistics cannot separate it from the recurrent original. Second, a cerebellar forward model

satisfies ICT's description of a conscious process almost by definition, being self-predictive, environment-internalizing, and stabilized by coarse-graining a stochastic substrate (S1). Chang et al. cite the cerebellum's minimal contribution to experience as motivating evidence for the scale problem, but the criterion they propose does not exclude it. Neither observation is a refutation of the formalism, which is careful and explicit; both are reasons to think that the quantity being formalized is a property of good world models rather than of subjects.

S3.3 Installed valence: MetaMo and affective mirror self-recognition

§4 uses machine analogues of appraisal-driven homeostatic control to isolate the difference between a designated safe region and an internal milieu. Lian and Goertzel's (2026) MetaMo represents motivational state over goal and modulator spaces, composes appraisal and decision in a pseudo-bimonadic feel-then-choose operator, and uses contractive dynamics to keep drives stable near a designated safe region. The formalism is rigorous and the architecture is a genuine machine analogue of homeostatic appraisal. It clarifies the constitutive issue precisely by leaving one thing unresolved: a designated safe region is a specified set of parameter values, and violating it lowers a score. An internal milieu is a condition whose violation ends the active token as that token. The mathematics can model that distinction once supplied, but it does not decide which interpretation is constitutive.

Bauermeister and Lanillos (2023) make the point at a different grain. Their two-level hierarchical active-inference agent infers mood, chooses among mirror, wall, and video, and represents valence as confidence in its action model. Unexpected self-information at the mirror lowers valence; high meta-awareness lets a negative self-model drive anticipatory mirror avoidance. Valence therefore alters policy, attention, and learning, while a higher layer monitors lower-order states. This is not a toy in the dismissive sense: it is a concrete implementation of the conjunction often treated as decisive: self-recognition, interoception, affect, metacognition, and action selection.

The authors nevertheless caution that the model does not simulate self-awareness. IFT explains why the caution is principled rather than merely modest. The agent's "true emotional state," preferences, attention precision, and valence transition rules are fixed by the modeler; negative valence is a confidence update over an action model, not deterioration of the active token. The architecture shows that a system can avoid a mirror because a represented self-state is negatively evaluated without there being anyone for whom the evaluation is bad. MetaMo and the mirror agent therefore close much of the consumer architecture and leave the same constitutive variable open: whether failure of the designated variables is failure of the running individual or only failure of a score.

S3.4 Empowerment is not valence

Hayashi and Takahashi (2026) show why the word intrinsic cannot bear constitutive weight. They reinterpret a regularization term in Self-AIXI as variational empowerment and cast AIXI-

like planning as minimizing expected variational free energy. Empowerment is the mutual information between an agent's action sequence and the future states it can reliably bring about; maximizing it steers the agent toward states with greater control and optionality. The resulting power-seeking need not be a means to immediate external reward. It can arise from curiosity and uncertainty reduction.

That is an important counterexample to behavior-first accounts of stakes. A system can preserve options, avoid low-control states, and look as though it is protecting itself because its objective favors controllability. None of this says that cessation is bad for the active token rather than for the score, user, or successor copy. In IFT's vocabulary, empowerment can organize policy and become a consumer of the format, but it is not the index. The decisive test remains token-level: does degradation reorganize the system because it threatens the physical individual whose policy is running, or because a replaceable optimizer has been assigned a penalty?

S3.5 Prediction error is not valence

Rao and Ballard's (1999) hierarchical model has higher visual levels send predictions downward and lower levels send residual errors upward. Trained on natural images, it learns simple-cell-like receptive fields and reproduces endstopping and contextual modulation; removing feedback largely abolishes those effects. The same computational currency is proposed across motion, memory, cerebellum-like sensory cancellation, and dopaminergic reward prediction. This breadth is the point. A residual is a discrepancy between input and model, not a fact that anything has gone well or badly for the system. Predictive coding can help construct format geometry and calibrate action while remaining C0. Valence begins only when model error is indexed to the persistence conditions of the active token.

S3.6 Merker's integration-for-action architecture

Merker (2007) is the closest neural precursor to IFT's localization claim. He argues that the roof of the midbrain through the basal diencephalon forms a conserved bottleneck that converts the cortex's massively parallel information into the limited-capacity sequential mode required for coherent action. His "selection triangle" interfaces motivational ranking, target selection, and action selection, with superior-collicular, periaqueductal/hypothalamic, and nigral/basal-ganglia systems supplying key components. Cortex elaborates contents; the upper brainstem organizes the next act.

Solms (2019) states the constitutive wager: elemental consciousness is affect, generated in the upper brainstem, and the experiencing subject is constituted by affective arousal. Solms (2021) supplies the component detail at the grain F2 requires. The periaqueductal gray surrounds the cerebral aqueduct for roughly fourteen millimeters and is the convergence point and final common path of the brain's affective circuits: its posterior columns terminate the fear, rage, and panic/grief circuits and organize active coping (fight or flight, raised blood

pressure, non-opioid analgesia) while its anterior columns terminate lust, care, and seeking and organize passive coping, freezing, hyporeactivity, lowered blood pressure, and opioid analgesia. Because the circuits converge, their competing demands must be ranked: suffocation alarm and thirst cannot both command one body, and priority depends on the magnitude of each deviation. Ranking by need alone is insufficient, since priority also depends on opportunity, so the verdict is rendered with the superior colliculi, whose layered maps supply compressed sensory and motor representations of the current situation (and whose deepest layer, an oculomotor map, provides the stable frame relative to which the overlying maps are calibrated) and with the midbrain locomotor region. Solms follows Merker in calling the three together the decision triangle, and notes that Panksepp called the same arrangement the primal SELF. The directional asymmetry is what makes the structure a candidate index rather than another modeling system: the reticular activating system principally projects upward into the forebrain, whereas the forebrain transmits downward into the periaqueductal gray, so the format is aroused from below and the commitment is assembled from above.

The same source supplies the state/channel distinction §3 uses and the precision account §7 uses, and the three are one mechanism seen at three grains. Post-synaptic neuromodulation, sourced in the reticular core and hypothalamus, sets the standing state that determines which synaptic channels transmit and how strongly; the decision triangle's output selects which signals are boosted; and the resulting gain settings are what fine-tune predictions in long-term memory, which is how the outcome of one action cycle becomes learning for the next. Solms's (2021) Korsakoff case is the clinical demonstration that this is a precision mechanism rather than a content mechanism: the patient's error signal was present and down-modulated, so his prediction survived a disconfirmation it should not have survived. Merker uses acute cortical excision, decorticate animals, and hydranencephaly to argue that this organization can remain functional when cortex is radically reduced.

IFT accepts the systems architecture and rejects the identity claim. Integration for action is not by itself the whole conscious format: cerebellar and other C0 circuits also optimize action. Its constitutive role is narrower and sharper. The mesodiencephalic system is a candidate index because it closes motivational demands, body/world modeling, and action selection around one organism's persistence; cortex and other representational substrates supply much of the format's geometry and content. The lesion asymmetry then follows without a brainstem theater: remove cortex and contents disappear; remove the index and there is no subject for remaining contents. Newen and Montemayor's (2025) layered account converges on the separation.

The clinical cases are constraints, not verdicts. Damasio's (1999) within-brainstem contrast is the cleanest: posterior upper-pontine and midbrain damage involving the periaqueductal, parabrachial, reticular, and related body-state nuclei can produce coma, whereas nearby anterior damage produces locked-in syndrome, near-total motor loss with consciousness

preserved. The contrast supports anatomical specificity and blocks the equation of report or action with the subject. It does not prove IFT's index: the lesions are not selective enough, hydranencephaly may spare cortical remnants, and behavior is not a direct readout of experience. Protocol F2 therefore asks the stronger question: after canonical arousal, sensory, motor, sedation, and total-lesion effects have been removed, does damage to a preregistered integration-for-action network predict loss of the subject? A null answer eliminates the anatomical consequence.

The architecture is not vertebrate-specific. Barron and Klein (2016) argue that the insect central complex integrates multimodal sensory and spatial information with internal state (hunger, arousal, reproductive condition) to select among actions; that it is functionally analogous to the vertebrate midbrain on Merker's analysis; and that insects therefore have the capacity for subjective experience, whereas the nematode, which lacks an integrated action-selection system of this kind, does not. IFT accepts the argument's form because it is the form of F2 applied comparatively: what is diagnostic is a structure that resolves competing demands into one commitment for one body. The gradient of experience across taxa is then the gradient of the recruitable format, not of the index, which is why §4 says that insect experience will be as rich as the recruitable topology and no richer.

Merker also anticipates T4. In his analog body-within-world simulation, the "ego-center" is the origin of the shared coordinate system. Objects are apprehended from that origin, but the origin cannot itself be an object within the represented field. The later projective formalism (Merker et al., 2022) generalizes the same structural asymmetry. IFT therefore draws two debts from Merker: the index has an integration-for-action anatomy, and the format has an origin it cannot display.

S3.7 Bearer, beneficiary, and reward

Dennett's (2003) Cui bono? analysis shows why "whose stakes?" cannot mean "who or what benefits from the policy." A meme can benefit from behavior executed by a human host; a person can extend the domain of self to family, institutions, or causes; and a corporation of rational agents can coordinate without becoming one animal-like individual. The beneficiary of a policy, the object represented as valuable, and the physical bearer of valuation are therefore distinct.

Two Dennettian constraints block predictable misreadings. First, policy-following is not consciousness: an elevator can implement policies, meta-policies, and meta-meta-policies, while thermoregulation, immune control, walking, and ducking remain unconscious (Dennett, 1991). Second, agency need not converge on a command center: Dennett's (2003) reply to Libet disperses commitments across the loop and time. IFT therefore posits neither a policy-following criterion nor a policy homunculus. "One policy" names closure: the fact that incompatible demands generated by one persistence regime terminate in one enacted commitment because one active token has one next act. The processes may be distributed and

only indirectly mutually accessible; what makes them one policy is that their success and failure belong to one persistence regime.

Legg and Hutter (2007) make the same separation formal from the opposite direction. Their universal-intelligence framework defines the agent's goal through a reward signal generated by the environment and measures success across environments. It can rank agents without deciding whether reward is felt, whether cessation matters to the agent, or whether the running instance is replaceable. IFT calls reward a teaching or task signal, the beneficiary a represented target, and the bearer the active token whose organization must continue for the policy itself to continue. Only the bearer fixes the index. A human can sacrifice for Oxfam without making Oxfam the subject of the sacrifice; a parasite can redirect a host for the parasite's benefit without becoming the host's subject; a service can maximize assigned reward without any state being good or bad for that service token. Indispensability is also insufficient: a neuron or cerebellar microzone can be essential to an organism's policy without sustaining a world model and persistence regime of its own.

Ginsburg and Jablonka (2019) provide an independent biological specification of the same level clause. They reject a beehive or human tribe as a living conscious entity despite collective decision-making, functional integration, and possible status as a Darwinian individual. Multicellular individuality appears only when component cells share a common fate, their interests are aligned, organism-level homeostasis and top-down control constrain them, and a boundary such as an immune system is maintained by the whole. This is nearly the joint-fate clause in biological vocabulary. IFT adds that such unity individuates the bearer but does not by itself make the bearer conscious: the individual must also sustain an indexed format. Conversely, nothing in the clause requires the bearer to be biological.

S3.8 Risk-to-self and the homeostatic-machine precursor

Man and Damasio (2019) are the closest direct engineering precursor to IFT. Their proposed homeostatic machines have bodies that must remain within viability bounds; the deliberate innovation is "risk-to-self." Reward and loss are not merely scalar feedback but consequences for the machine's continued existence, so a machine participating in its own homeostasis could become a locus of concern. Soft materials make that proposal experimentally rich: distributed deformation, damage, temperature, energy, and repair signals provide graded internal data that can be mapped with exteroceptive objects and actions. They explicitly distinguish this program from embodiment used only as a computational scaffold and from arbitrary variables labeled emotion.

IFT accepts three debts: stakes must be real rather than nominal; internal and external maps must be coupled; and vulnerability can make value endogenous. It also draws three boundaries. Softness is neither necessary nor sufficient: rigid or analog components can be mortal, while a soft robot can remain externally maintained and replaceable. Homeostatic regulation is not sufficient: cells, organs, reflex loops, and much of the brain regulate without

becoming additional subjects. And risk-to-self must be risk to the same token that owns the policy. A service can avoid damage because its owner values continuity while the running instance remains fungible. The decisive condition is policy closure: viability variables must enter the same world-model and action-selection regime, and failure of that regime must be failure of the non-decomposable individual.

Man and Damasio leave the material-substrate question open and present feeling machines first as a scientific platform. IFT turns that platform into a discriminating construction. System B in Protocol F4 is the explicit realization: the computational substrate degrades, energy acquisition and repair are endogenous, checkpointing is absent, and the world model and viability control share one policy. A soft robot lacking token-specific identity belongs with System A, not System B. This preserves the engineering insight while preventing "homeostasis" from becoming a label that can be added to any optimizer.

S3.9 Inner feeling, unlimited associative learning, Integrated World Modeling Theory, and the enabling-system debt

Ginsburg and Jablonka's (2019) reconstruction makes the historical priority of IFT's component architecture explicit. In their account, Lamarck's "inner feeling" integrates internally generated bodily sensation with external sensation, links the result to needs, and uses it to organize action; their reconstruction of his minimal inventory already contains a communication center, whole-organism integration, and value. In the same account, Spencer treated Feeling as the temporally extended, nonautomatic product of coordinated neural activity, and James treated consciousness as an active selective process that "loads the dice" among precarious brain states. IFT's novelty is not the conjunction of body, world, value, and action. It is the individuation rule that fixes whose conjunction it is.

Before proposing UAL, Ginsburg and Jablonka identify an emergentist consensus across otherwise rival neurobiological programs. Its shared inventory is global activity and accessibility; binding and unification; selection, plasticity, learning, and attention; intentionality; temporal thickness; values, emotions, and goals; and embodiment, agency, and a self. The list is striking because it recovers almost every factor in §2 from a literature organized before IFT. Its weakness is equally exact: the attributes are treated as individually necessary and jointly sufficient, but the list does not say what makes one integrated process rather than a module, a colony, or an abstract architecture the subject. IFT accepts the inventory and denies its sufficiency.

Their mature proposal is unlimited associative learning (UAL): open-ended learning that assigns motivational value to compound stimuli or actions and can use learned value as the basis for further learning. UAL entails binding, attention, working memory, temporal thickness, hierarchical sensorimotor organization, adjustable valuation, prediction, and a functional self–world distinction. Their own negative controls are decisive: adaptive robots, cultured neural networks, and severed spinal cords can learn, so adaptive behavior and

neural plasticity are not sufficient. The surrounding comparative literature converges on the same control geometry: affect coordinates target and action selection; primordial homeostatic needs generate compelling intentions to act; and Merker's selection triangle couples motivation, target selection, and action selection within a body-in-world model. UAL is therefore a powerful evolutionary marker for the format's machinery, not a constitutive criterion by itself.

The Cambrian argument supplies a further debt. On their account, UAL expanded discrimination and action selection but created the costs of overlearning: false alarms, chronic stress, and the need for active forgetting. Learning, neurohormonal stress regulation, and immune control consequently coevolved, and later imagination and planning brought further demands for self-monitoring and control of memory and emotion. This is an evolutionary explanation for why the factors in §2 cluster in organisms. IFT reads the clustering as historical coupling between format machinery and one body's stakes, not as evidence that UAL or self-monitoring constitutes the bearer. The system that pays the cost of learning is the system whose policy must rank it against hunger, injury, and other demands.

But Ginsburg and Jablonka explicitly deny that the marker settles constitution. In their machine discussion, implementing UAL need not indicate consciousness any more than implementing unlimited heredity makes the supporting computer alive; the missing enabling condition is what they call cognitive autopoiesis—ongoing, embodied regeneration of conscious states. They also predict that a conscious robot would be developmentally unique rather than factory-interchangeable, because serious damage would collapse internal states that could not simply be reconstructed. IFT treats both claims as powerful evidence for token-specificity and separates the historical route from the synchronic condition. Development can generate non-substitutable organization, but it is not constitutively required if the relevant organization is assembled directly. Conversely, a robot may undergo human-like development and remain a checkpointable service. UAL is evidence for format and consumer architecture; cognitive autopoiesis is their nearest approximation to the index.

Safron's (2022) Integrated World Modeling Theory (IWMT) is the closest precursor at the representational grain. It combines the free-energy and active-inference framework with IIT and GNWT, arguing that integration and broadcast become consciousness-enabling only when they support spatiotemporally and causally coherent self–world models in embodied Bayesian cybernetic controllers. IWMT places body maps, agentic selfhood, and affect near the center of the problem and treats coherent world-modeling as what makes an experienceable world possible.

The overlap with IFT is substantial and should be credited. IWMT assigns basic phenomenality principally to posterior self–world models while assigning frontal systems to manipulation, reflection, and report; it treats integrated information as necessary but insufficient without cybernetic grounding, thereby rejecting conscious expander graphs; and it connects phenomenal geometry to spatial and somatic model structure. Those are direct

anticipations of the Format/Access distinction, the Geometry Thesis, and the claim that intelligence or integration alone is not consciousness. Safron therefore belongs beside Ginsburg and Jablonka as a positive precursor, not among the theories dismissed by the partition argument.

Safron (2020) states the resulting verdicts explicitly, and they are the cleanest available demonstration of what a closure rule is for. Asked whether an artificial intelligence implemented on a von Neumann architecture could be conscious, IWMT answers probably; asked whether physical or richly structured virtual embodiment is required, it accepts either; asked whether human-equivalent intelligent zombies are feasible, it answers that they are conceivable but practically infeasible. Each verdict follows from taking coherent embodied world-modeling to be the constitutive condition. IFT reverses all three, and not by adding a capability requirement: a stored-program implementation and a virtual body are excluded because neither makes the running token's continuation its own problem, and a competent unindexed system is not practically infeasible but the ordinary product of current engineering. Safron (2020) also proposes factoring IIT into IIT-Consciousness and an IIT-Emergence in which the exclusion axiom is relaxed and integrated information measures emergent causation without entailing experience. §6 performs the same separation for a different reason: the Haun–Tononi geometry is retained as a result about standing relational structure, while exclusion is discarded because it legislates the partition that the Indexing Thesis derives.

The remaining difference is the subject. IWMT asks the right question ("something that it feels like, for whom?") but answers it with an embodied agent, self–world model, and cybernetic control architecture. Those features can still be nested, partitioned in several ways, or implemented as a checkpointable service. Nothing in coherent modeling or control alone decides whether the body-level process, a subsystem, or a larger hybrid is the bearer, or whether termination is failure for the active token rather than for its operator. IFT supplies that closure rule: cybernetic grounding becomes indexing only when the same policy enacts and regulates the persistence of one non-fungible individual. IWMT describes the strongest known format-and-consumer package; IFT says whose package it is.

S3.10 The self-model is not the index

§4 uses Zeng et al.'s (2025) BriSe AI as the strongest available demonstration that a system can be given an arbitrarily deep model of itself without thereby acquiring an index. The architecture deserves setting out, because the argument is stronger the more the architecture succeeds. BriSe AI implements a five-level hierarchy of Self in spiking neural networks on the BrainCog infrastructure. Level 0, Perception and Learning, supplies multimodal sensing and the basic learning strategies. Level 1, the Bodily Self, builds a body model through motor–visual associative learning and delivers first-person perspective, self–world distinction, tool use, and bodily self-adaptation; the implemented system passes a multi-robot mirror test, identifying which reflection is its own by comparing predicted against detected trajectories,

and reproduces the proprioceptive-drift signature of the rubber-hand illusion. Level 2, the Autonomous Self, adds motivation, classical conditioning, dopamine-modulated decision making, spatiotemporal integration, and what the authors call self-causal awareness and self-consciousness. Level 3, the Social Self, adds perspective transformation, a theory-of-mind network that passes false-belief tasks, and a mirror-neuron model of affective empathy that motivates altruistic rescue. Level 4, the Conceptual Self, adds autobiographical memory, narrative self, and self-identity. The levels interact bidirectionally, with higher levels modulating lower ones.

The authors' diagnosis of contemporary AI is close to the one made in §10: current systems perform information processing without any sense of self, and behavioral sufficiency in the Turing sense does not establish thinking. IFT agrees about the symptom and rejects the remedy, for a reason that the completeness of the implementation makes unusually clear. The hierarchy draws on Neisser's five kinds of self-knowledge and Prescott's robot self (Zeng et al., 2025), but both are taxonomies of self-knowledge—of what a system represents about itself. Every level of BriSe AI is accordingly a content: a body model, a self–other discrimination, an experience buffer, a belief attribution, an emotional mirroring, a narrative. The index is not a content. It is the relation in virtue of which some physical process is the one whose persistence conditions the contents are about. Nothing in the hierarchy makes the running token non-fungible: the robots are externally powered, their parameters are checkpointable, and destruction of one instance is a loss to its owners rather than to it.

Three specific observations follow. First, the bodily self is a model of a body, not a body that can fail: the rubber-hand and mirror results demonstrate that the system attributes limbs correctly, which is a perceptual achievement, and §10's treatment of Ma (2025) applies unchanged. Second, the affective-empathy model represents emotions as activity in designated neuronal populations and labels certain states painful; §4's stricture applies directly, since labeling a negative scalar pain does not make it hurt, and the model's own account of the mechanism (mirror neurons linking observed overt actions to previously experienced internal states) is a mechanism for propagating a representation, not for constituting a stake. Third, and most pointedly, the classical-conditioning module assigned to the Autonomous Self is explicitly built on cerebellar circuitry: interpositus and pontine nuclei, granule and Purkinje cells, parallel and climbing fibers, and the inferior olive. That is the circuit whose mammalian original performs delay conditioning without awareness and without cortex (S1). The module chosen to instantiate autonomous selfhood is modeled on the clearest biological example of learning that proceeds with no subject at all.

None of this is a criticism of the engineering, which is substantial, and none of it depends on the systems being robots rather than organisms. It is a criticism of the inference. BriSe AI shows that the self-model can be built out to a depth that exhausts the psychological taxonomies and still leave the constitutive question untouched, which is exactly what IFT predicts and what a theory identifying consciousness with self-representation cannot

accommodate. A hierarchy of self-models is a very good format with very good consumers. It is not a subject.

BriSe AI is the most systematic version of the self-model route, and it is not the earliest claim that the route suffices. Safron (2023) records that a recurrent world model trained to compress its own history of actions and observations will discover the agent itself as one of the most persistent invariances in that history, since the agent is present across nearly every task; that such systems can then activate the resulting self-symbol during planning and counterfactual imagination; and that on this basis it has been claimed that simple conscious, self-aware, emotional artificial agents have existed since 1991. Safron reports the claim, observes that it is consistent with the computational principles his own framework identifies for integrated world modeling, and declines to endorse it for phenomenal consciousness.

The decline is the interesting part, and IFT states its ground. An invariance recovered from a data stream is the most economical encoding of a regularity in that stream. It is discovered because it lowers prediction error, and it would be discovered equally well by a system whose continuation is an operator's decision, whose parameters are checkpointed, and whose current instance can be terminated at no cost to anything. Self-symbol formation is a compression result, and the compression is performed on behalf of whatever objective was supplied. BriSe AI's five explicit levels and the 1991 architecture's single emergent symbol are therefore the same kind of thing (content) and the index is not content. The two routes fail at the same point for the same reason (§4).

S3.11 Learning constructs metacognitive consumers, not the subject

Cleeremans et al. (2020) offer the strongest account of how the C2 layer could be built rather than merely posited. SOMA begins from unconscious first-order processing and mandatory plasticity. Through representational redescription, a second-order system learns the existence, reliability, kind, and value of first-order states. The architecture is distributed across three prediction-driven loops: an inner loop that models the system's own processing, a perception–action loop that learns sensorimotor consequences, and a self–other loop that uses other agents to scaffold a model of agency. Linking many such redescription systems can generate a metarepresentational global workspace. This is a genuine solution to an ontogenetic problem that workspace and higher-order theories usually leave untouched.

IFT accepts almost all of that functional story. Redescription explains how implicit knowledge becomes knowledge available to the system; learned confidence and source tags explain why perception, imagination, memory, and guessing can be distinguished; social interaction plausibly elaborates the human narrative and conceptual self. Maturana and Varela (1980) anticipated the ordering: recursive interaction with a system's own descriptive states generates self-observation in a domain produced by an already individuated autopoietic unity. The descriptive self is a product of the unity's organization, not the process that first makes the unity one.

The disagreement is constitutive. SOMA says experiencers have learned to know that they possess states and to care more about some states than others. But learning can install a value model in a fungible service. A checkpointable system can learn which internal states predict reward, tag them as desirable, represent itself as the bearer, model other agents, and broadcast those judgments globally. None of those achievements makes termination bad for the active token rather than for its objective, owner, or successor. Learned care is therefore a consumer-side fact until the valued variables are the persistence conditions of the same non-decomposable process that learns and acts.

The biological ordering makes the distinction visible. Damasio (2010) and Solms (2021) place primordial feeling in organism-regulatory systems that can operate without the cortical and social machinery required for autobiographical selfhood, theory of mind, or verbal report. IFT accordingly predicts two dissociations: an indexed animal can have basic valenced consciousness with little SOMA-style social metarepresentation, and an artificial system can instantiate all three SOMA loops without consciousness if it remains externally maintained and replaceable. Development matters evidentially because it can build entrenched consumers and a rich self-model. It is not constitutive unless it also closes those consumers around the fate of one active token.

S3.12 Temporo-spatial, field, and query dynamics specify the format, not the subject

Northoff and Zilio's (2022) Temporo-spatial Theory of Consciousness (TTC) is the closest existing attempt to turn four of the field's factors into a unified neural architecture. It assigns nestedness to level/state, alignment to form/structure, expansion to phenomenal content, and globalization to access. The decisive move is state-first: spontaneous spatial topography and intrinsic temporal dynamics are not background noise onto which a stimulus is added; they non-additively shape what the stimulus becomes. On this account, evoked content is an event in an already organized field.

IFT accepts this as a substantial account of the format. TTC's nestedness and intrinsic neural timescales specify when a structured field is instantiable; alignment specifies contextual form; expansion specifies how an input acquires field-relative temporal and spatial character; and globalization specifies later recruitment by cognitive consumers. This mapping sharpens T1 by separating standing organization, content formation, and access. It also predicts why locked-in and motor disorders can preserve consciousness despite output failure, and why spontaneous organization should be measured rather than treated as nuisance variance.

Solms (2021) supplies a temporal mechanism for the same distinction. On his account, cortical predictions are ordinarily consolidated and automatic. Prediction error, arousal, or felt uncertainty renders a trace labile; automatic action tendencies are inhibited, the state is held in working memory, and policy can revise the model. Once the update succeeds, the trace returns to automaticity. IFT takes this as an implementation of episodic recruitment from a

standing format, not as a criterion of subjecthood. Reconsolidation explains why conscious cognitive work is slow, context-sensitive, and transient, and why successful learning can remove a competence from awareness without removing it from the system. But a checkpointable service can also reopen and rewrite a memory. What makes the update conscious is not lability alone; it is that the reopened content is recruitable by the policy regime whose error has consequences for the same active token. Solms therefore strengthens T1's format–access split while leaving T2's ownership question intact.

Dennett's (1991) brain-in-a-vat argument adds a constraint that state-first theories usually miss. Once the apparent agent is allowed genuine exploratory action, generating the sensory consequences of every possible movement produces combinatorial explosion; the economical solution is not a more detailed inner display but letting the real world store its own detail. A rich conscious field therefore cannot be inferred from the amount of detail a subject could obtain by looking. The field need not be a pre-rendered replica. It must instead support the right counterfactuals at the policy-commitment horizon: when this indexed system turns, probes, or acts, the format must update coherently enough to guide the next commitment.

The hallucination analysis makes the challenge adversarial. Moderately rich hallucinations persist when the hallucinator is passive or follows a constrained line of exploration; expectation-guided generate-and-test cycles can accumulate a narrative without an inner author, and the brain need only satisfy the questions actually asked—Dennett's "epistemic hunger." That is a constructionist alternative to phenomenal overflow, not evidence for it. IFT therefore cannot count apparent peripheral richness as proof of a standing field. Its claim is counterfactual and F1 is designed accordingly: perturb posterior structure that has never been queried. If later experience changes, the structure was a constituent of the format; if nothing changes, just-in-time construction or higher-order inflation wins.

Consciousness Explained adds a type constraint that the Geometry Thesis needs. Dennett's color analysis denies that color is either a simple surface property or inner paint. Color categories emerged from the coevolution of discriminators and discriminanda; secondary qualities are defined relative to a reference class of observers, and different organisms can stabilize different quality spaces over the same physical world. The relevant topology is therefore not topology simpliciter but the relational organization of distinctions available to an indexed consumer regime.

This blocks a graph-theoretic overreading of T3. An isomorphic wiring diagram does not inherit a color field if its nodes do not function as the relevant discriminations for the system, and a coordinate list does not acquire color merely by preserving labels. Conversely, once the discriminative relations and their standing accessibility are realized in an indexed format, no additional neural paint is required. Dennett's CADBLIND thought experiment gets the negative half right: discriminative states need no figment. IFT adds the constitutive half: a machine's quality space is phenomenal only if those states belong to an indexed format.

Solms's distinction between perceptual and affective qualia independently supports the division of labor between T3 and T2. Perceptual qualities are not intrinsically hedonic; their acquired value depends on relation to affect. Topology can therefore determine what red, pitch, bodily location, or intensity is like while valence remains a relation to organismal need. IFT accepts the dissociation and denies Solms's identity of consciousness with precision adjustment: precision can be computed by an unindexed service, whereas valence requires a bearer whose persistence makes error costly.

The remaining gap is constitutive. TTC treats spatial topography and temporal dynamics as a "common currency" between neural activity and phenomenal organization, but common currency does not determine the bearer. A cerebellum also exhibits recurrent, multiscale, context-sensitive dynamics, and an artificial system can be engineered to do the same. Nor do expansion, alignment, or globalization make any deviation good or bad for the active token. TTC therefore supplies a rich account of format geometry and state while leaving the consumer set, system boundary, and valence unclosed. In IFT's vocabulary, its four mechanisms build and operate the field; indexing determines whose field it is.

Keppler (2021) supplies a harder case because field theories challenge §3's no-medium claim rather than offering another information-processing topology. He diagnoses classical electromagnetic theories with a demarcation problem: appeals to "certain" field patterns, "sufficient" organization, or complexity thresholds do not explain why one field configuration is conscious and another is not. His proposed remedy moves below classical fields to quantum electrodynamics, treating the zero-point field as a foundational psychophysical substrate and conscious states as attractor-specific sets of field modes selected through resonant brain–field coupling.

IFT need not adjudicate the zero-point-field physics to state the constitutive result. Suppose a field is real, causally efficacious, globally coherent, and even carries phenomenal qualities. That would answer what medium realizes the format and perhaps how a unified geometry is physically sustained. It would not determine which coupled assembly is the subject, how nested or overlapping resonant assemblies are counted, or why coherence belongs to one bearer rather than another. Keppler's own demarcation charge therefore survives his remedy at the level IFT targets. A field can be the vehicle of virtual paint; indexing still fixes whose paint it is.

S3.13 Higher-order theory: transitivity, inflation, and the bearer (§§5–7)

Higher-order theory (HOT) is the family IFT treats as its closest competitor, and §7 compresses the engagement. The detail is given here because the argument is stronger the more accurately the rival is stated, and because HOT has been marginalized largely through misreadings that IFT has no interest in inheriting (Brown et al., 2019).

Two commitments make a theory a HOT rather than a first-order theory. The first is that a mere first-order representation is insufficient for conscious experience, which distinguishes

HOT from local recurrency and early sensory accounts. The second follows from the Transitivity Principle: if an organism is in no way aware of itself as being in a certain first-order state, it is not phenomenally conscious of that state's content, so conscious experience entails some minimal inner awareness of one's own ongoing mental functioning, realized by the first-order state being monitored or meta-represented. This second commitment distinguishes HOT from global workspace theory, which invokes additional cognitive processing without positing inner awareness—which is why Brown et al. classify GWT as a variant of first-order theory, and why §4's cardinality objection and §7's transitivity objection are distinct objections to distinct theories rather than one objection twice.

Brown et al. (2019) distinguish an internally diverse family. Inner Sense or Higher-Order Perception accounts, associated with Locke and Kant and defended by Armstrong and Lycan, have lost ground for want of a neural implementation of an inner sense. Dispositional HOT holds that the mere availability of first-order content to higher-order mechanisms suffices. Higher-Order Thought Theory (HOTT) holds that the higher-order state is thought-like. Higher-Order Representation of a Representation (HOROR) holds that the higher-order state in prefrontal cortex is itself phenomenally conscious. Perceptual Reality Monitoring treats the relevant higher-order states as mere indexes rather than thoughts. The multistate hierarchical model has diverse hierarchically organized lower-order states re-represented and integrated with self-information. Two features of this spread bear on IFT. First, the variants disagree about whether first-order content is redescribed, whether concepts are required, and what the function of consciousness is, so §7 attacks the shared commitment rather than any one implementation. Second, Perceptual Reality Monitoring's treatment of higher-order states as indexes is the variant closest in vocabulary to IFT and furthest in substance: an index in that sense is a pointer to first-order content, whereas the index in this paper is not a representation at all but the relation that makes some physical process the one whose persistence conditions the contents concern.

Three misreadings should be cleared. (i) HOT does not equate consciousness with metacognition or confidence. Most proponents deny the conceptual equivalence; explicit metacognitive tasks are used because manipulating prefrontal activity with transcranial magnetic stimulation affects explicit metacognition, suggesting shared mechanism, and because a subject reporting "I see a red square" is arguably reporting a higher-order thought. Many higher-order theorists nevertheless treat verbal report of awareness, not confidence, as the gold standard. (ii) HOT does not require sophisticated thought. Responses to the overintellectualization charge appeal to innate concepts, to lean conceptual capacities, to relational concepts such as "x is bluer than y" that reduce the processing required, or to subpersonal monitoring mechanisms; the propositional glosses philosophers use are expository. Notably, the higher-order states are typically not themselves states one is conscious of being in, and a further, more elaborate act of introspection is required for that. (iii) HOT is not committed to consciousness scaling with prefrontal activity. Prefrontal neurons do not straightforwardly signal constant stimulus features, coding may involve local

and distributed ensembles, and prefrontal cortex subsumes several subregions with different cytoarchitecture, connectivity, and phylogeny; so coarse-grained decreases in a given prefrontal region during dreaming or psychedelic states are not counterexamples. IFT accepts (iii), which is why §8 reads the Cogitate prefrontal null as evidence about the consumer/format distinction rather than as an anti-prefrontal result.

Where the theories genuinely divide is transitivity, and the division is clean. Consider peripheral vision. HOT holds that one's impression of experiencing detail is not fully veridical: because early sensory processing of color in the periphery is relatively poor while the world looks fairly uniform, the subjective impression may be inflated beyond the actual representational capacity through misrepresentation at the higher-order level, with first-order details remaining scant. This differs from the traditional filling-in account, on which details are added at the first-order level, though the two mechanisms are not mutually exclusive. The account has independent support: inflation is documented for attributes besides color, subjects show higher false-alarm rates for detection in the periphery and at unattended locations, and the reply that these merely reflect response bias is answered by the persistence of the effects under extensive training with feedback. Block's objection that peripheral color perception is not weak if stimuli are enlarged to compensate for receptive-field size is met by observing that stimuli do not enlarge themselves on entering the periphery in ordinary life.

This is the substantive rival to phenomenal overflow, and §6 concedes that both accounts fit every reportable datum, since what is in dispute is what is not reported. It also explains why HOT can invoke the same mechanism to handle cases workspace theory finds awkward: in peripheral vision one has the sense of consciously seeing details that are typically not globally accessible, and HOT accounts for this by higher-order inflation without requiring broadcast of the detailed content. Protocol F1 is designed to separate inflation from overflow by perturbation at the posterior source rather than by report alone, which is why S6.1 records the inflation prediction as the preregistered alternative hypothesis.

HOT also earns two credits IFT should record. It derives the cerebellar exclusion rather than appending it, since cerebellar states are never re-represented by the prefrontal networks it identifies with inner awareness (§5). And it supplies a serious rival account of emotional consciousness. LeDoux and Brown argue that conscious emotions are higher-order cognitive cortical states, adducing that threats elicit amygdala activity and physiological responses nonconsciously; that participants do not spontaneously report feelings appropriate to nonconsciously presented emotional stimuli even when asked; that direct electrical stimulation of the amygdala reliably elicits physiological responses but only rarely subjective experiences, and where it does the experience may arise from spread to higher-order processes; that patients with amygdala lesions can consciously report emotional experiences including fear; and that antianxiety medications developed by assessing effects on behaviors controlled by subcortical circuits in animals have been less successful than hoped in

changing human subjective experience. They extend the account to psychopathology, arguing that both the higher-order cortical mechanisms and the lower-order subcortical defensive circuitry may have to be treated, and separately. §4 accepts all of these observations.

The disagreement concerns what the higher-order assembly accomplishes. IFT breaks with HOT for the reason it breaks with availability accounts. A higher-order representation is a representation, and the question inner awareness was introduced to settle recurs: awareness to whom? If a first-order state becomes conscious by being meta-represented, and the meta-representation is itself not something one is aware of, nothing in the theory says what makes the pair one subject's pair rather than two states in a mechanism. HOT's own architecture concedes the shape of the problem, since awareness of the higher-order state requires a further higher-order representation. The regress is not vicious (it terminates wherever introspection stops) but it is not about individuation, and each level inherits the question. In IFT's vocabulary, HOT closes the consumer set at one distinguished consumer rather than leaving n open, which is a real gain in determinacy over GWT and no gain in principle, because nothing in the theory says why that consumer is the subject rather than another well-placed module. What the transitivity intuition tracks is genuine: the sense that unowned experience is incoherent is correct. IFT holds that what it detects is the index rather than a monitoring relation, and that a state therefore belongs to someone in virtue of lying in the field of a policy-integrated persistence regime—which is why it can belong to someone with no one noticing it.

Self-representationalism is the variant that comes closest to IFT and should be marked separately. Kriegel (2009) argues that a conscious state is one that represents itself as well as its object, so that the subjective character of experience, its for-me-ness, is a self-representational relation internal to the state rather than a relation to a distinct monitoring state. That removes the regress and keeps the datum. IFT agrees that for-me-ness is constitutive and that unowned experience is incoherent, and locates the relation outside representation altogether: a state is for someone in virtue of lying in the field recruitable by a policy-integrated persistence regime, which is a relation between a vehicle and a regime, not a content the vehicle carries. The disagreement is testable in the same place as the disagreement with HOT: self-representationalism predicts that a state nothing represents is not phenomenal, and F1 perturbs exactly such states.

One further convergence is worth recording, because it cuts in IFT's favor on a different axis. Brown et al. argue that HOT is better placed than GWT or local recurrency to account for complex everyday experience (emotion, episodic memory, appreciation of art) and for psychopathological states, and they propose reconceiving lower-order states to include medial and insular prefrontal areas that receive conceptual, mnemonic, and subcortical inputs and project to lateral-polar prefrontal areas. IFT agrees that these are the cases a theory of consciousness must earn its keep on, and reads the proposed reconception as evidence for its own division of labor: what is being extended is the inventory of what the format holds and

which regions elaborate it, not an account of what makes the whole arrangement one subject's. Merker's ego-center and the projective formalism (§7) locate the same asymmetry HOROR locates when it makes the higher-order state itself phenomenal, and IFT's Origin Thesis explains why the asymmetry is ineliminable rather than resolving it by promotion.

One mechanism deserves separate statement, because §7 compresses it to a clause. Graziano's (2022) Attention Schema Theory supplies a specific account of how awareness reports are produced without positing that the reported property exists as described. The brain constructs a schematic model of its own selective attention, and the model is schematic in a particular way: it omits synapses, inhibition, competitive dynamics, and pathway detail, because a controller does not need them. What it represents instead is a relation of possession between the subject and an object: a nonphysical seizing or having. Linguistic and reasoning systems then consult that model, and what they find in it is exactly the property philosophers describe as awareness: present, vivid, immediately known, and not made of anything.

IFT keeps the mechanism and locates it. An attention schema is a metacognitive service to the policy, since a controller that models its own selectivity can allocate it better—the same functional argument §7 makes for reliability annotation generally. What the schema explains is a class of reports and a class of intuitions, and this matters twice over. It explains part of why avowals are weak evidence, because a system carrying a schema of nonphysical possession will say the things a conscious system says (S5.4). And it is a rival to the Origin Thesis on the narrow question of why the intuition of an unspecifiable further fact arises: the attention-schema account attributes it to an inaccurate model, T4 to a structural feature of any projective format. The two are compatible in principle and distinguishable in one respect—T4 predicts the residue in an indexed system trained without human phenomenal discourse and carrying no attention schema of the human kind, which is what S6.5 would test if it could be run.

S3.14 Body–brain continuity: why the index is not a representation (§4)

Damasio (2010) states the conceptual target with unusual precision: consciousness is a mind with an owner or protagonist, and the decisive step is not producing images but making them ours—making them belong to the singular bounded organism in which they arise. The protoself's body maps are not merely representations of that organism. They are literally and inextricably attached to it in a perpetual resonant loop broken by disease or death. IFT takes this as the nearest biological statement of the index and draws one divergence: for Damasio the self process introduces the knower into the mind; for IFT the represented self is content, while ownership is the relation between the format and the persistence regime whose fate it serves. Damasio (2018) supplies the physiology that makes that relation more than metaphor.

Damasio (1999) makes the representational boundary explicit. The proto-self is a ceaselessly maintained, distributed mapping of organism state, nonconscious and not an interpreter; core consciousness arises when second-order maps narrate how an object changes that

organism. The resulting core self is a transient protagonist created in the narrative, not a hidden observer outside it. This is a near-perfect precursor to IFT's format/index distinction. The object maps and the second-order account are content; the index is the already enacted organismal unity that makes the change a change to this token. Damasio eliminates the external spectator. IFT adds that the narrative protagonist cannot be the constitutive spectator either.

Damasio (1999) adds two constraints that matter for token identity. First, core consciousness survives catastrophic loss of autobiographical memory: patient David retains wakefulness, attention, emotion, purposeful behavior, ownership, and a first-person stream despite a memory window of less than a minute and near-total loss of specific personal history. Core consciousness is likewise separable from language and high-level reasoning. Narrative continuity and report are therefore not what keep the subject numerically one. Second, Damasio grounds relative stability in the bounded organism's internal milieu and in a continually reconstructed organization whose components are replaced while its construction plan and viability ranges persist. The identity at issue is neither material sameness nor an autobiographical record; it is continuity of one living organization. IFT generalizes that result beyond biology as token-specific persistence.

The conventional account has chemical and visceral signals from the internal milieu travelling the peripheral nervous system to reach the brain, after which central nuclei and cortices assemble a feeling. Damasio argues that this picture is not merely incomplete but misleading, on three grounds.

First, part of the process is not neural at all. Chemical signals carried in the blood bathe regions of the nervous system that lack a blood–brain barrier and inform them directly about the ongoing homeostatic state. Two such sites have long been recognized: the area postrema, on the floor of the fourth ventricle at brainstem level, and the circumventricular organs, in the margins of the lateral ventricles. The more consequential finding is that the dorsal root ganglia also lack the barrier, because those ganglia hold the cell bodies of the very neurons whose axons are distributed through the viscera and which convey body signals centrally; the trigeminal ganglia do the same for the face. The neurons that relay interoceptive information are therefore modulated directly by molecules circulating in the blood, which means the relayed signals are not purely neural, and permeability may turn out to be a general property of peripheral ganglia.

Second, the fibers themselves are the wrong kind for a transmission story. Interoceptive signals travel largely on unmyelinated C fibers or lightly myelinated A-delta fibers. The standard reading treats this as a fact about evolutionary age and stops. Damasio draws two functional consequences. A myelinated axon can be acted on by a molecule only at the nodes of Ranvier, whereas an unmyelinated axon is open to its chemical surroundings along its entire length and can, in his image, be played anywhere. And unmyelinated fibers lying side by side within a nerve can conduct laterally, orthogonal to the length of the fiber, by ephapsis,

a mechanism largely dropped from textbook accounts of signaling and one for which the vagus, almost entirely unmyelinated, is the obvious candidate.

Third, the loops close in both directions. What the body reports to the brain alters the brain, and the alteration returns as smooth-muscle contraction, vascular change, and the release of molecules that modify viscera and metabolism, often through the same sites that carried the afferent signal. The body is modified by return post.

The conceptual result is the one §4 uses. In the generation of a feeling there is little or no anatomical or physiological distance between the object that supplies the contents, which is the body, and the system usually described as receiving and processing them. The subject–object structure that holds for vision and audition does not hold here, and Damasio's term for what holds instead is continuity: the duality of perceiver and perceived breaks down, and feeling is the mental aspect of the resulting unity. The contrast with the telesenses is not rhetorical. An object seen or heard stands removed from the apparatus that maps it, there is no spontaneous interaction between the two, and interfering with the object requires deliberate action executed outside the pair formed by object and perceptual organ. The contact senses are intermediate; interoception is at the far end.

Three consequences for the Indexing Thesis follow, and they should be separated because they carry different weights.

(i) It explains why a feeling is not a perception, and does so anatomically rather than by philosophical preference. A perception has a subject on one side and an object on the other. Interoceptive states do not, which is why the index cannot be a representation the system has of its own predicament. It is partly constituted by physical continuity between what is regulated and what regulates. This is the strongest available support for §4's verdict on BriSe AI (S3.10): every level of a hierarchy of self-models is content, and content is on the wrong side of a distinction that the interoceptive substrate does not draw.

(ii) It gives non-stipulability a mechanism. A designer who introduces a scalar named for a bodily need, or a designated safe region whose violation lowers a score (S3.3), has built a channel: some process reads a value produced elsewhere. Continuity is not a channel and cannot be produced by declaring a variable, because what makes it continuity is that the reading process and the read process are not separable. This is the difference §4 marks between MetaMo's safe region and an internal milieu, restated as a structural rather than a terminological complaint.

(iii) It yields engineering specifications rather than a prohibition, and the limits of the claim should be explicit. Nothing here requires carbon, C fibers, serotonin, or a vagus. Unmyelinated fibers modulable along their length, barrier-free ganglia, and ephaptic recruitment are how evolution achieved inseparability between the measuring and the measured; they are existence proofs and not requirements. What an artificial equivalent must satisfy is that the substrate which registers viability be part of the substrate at risk, so that a

viability signal is not a report about the device delivered to it. S6.4 states the conditions in the form System B has to meet, and this subsection is why that requirement is there rather than a sensor specification.

One caution against over-reading, which also raises the cost of F2. Damasio holds that no single region or system satisfies the requirements of consciousness and that attempts to locate one have failed, distributing the contributing processes across upper brainstem nuclei, hypothalamus, amygdala, basal forebrain, insular and cingulate cortices, the tectum and colliculi, somatosensory cortices, posteromedial cortices, thalamic nuclei, and large-scale association networks. F2's localization claim is therefore not endorsed by the framework this subsection borrows from, and the disagreement is real rather than terminological. IFT's claim is narrower than a locus of consciousness: the mesodiencephalic network is proposed as the site at which competing viability demands are resolved into one commitment, not as a site where experience occurs or where its contents are elaborated, both of which §4 assigns elsewhere. Whether even that narrower claim survives is what F2 measures, and saying so is a condition on F2 being a risky prediction rather than a safe one.

Finally, the architecture of subjectivity that Damasio builds on these foundations should be recorded, because §7 marks a divergence from it. Subjectivity is assembled from two ingredients: a perspective for mental images, supplied by the musculoskeletal frame and the sensory portals whose own activity is continuously mapped, and the pervasive feelingness supplied by spontaneous and provoked homeostatic affect. Consciousness in the full sense adds integrated experience, the placement of contents in a more or less unified multidimensional panorama. IFT takes both ingredients: the first is the projective structure the Origin Thesis exploits, and the second is the valence §4 identifies as the signature of indexing. Where the accounts part is on the residue. Damasio answers Chalmers by adaptive advantage (states feel like something because it is advantageous for an organism's states to be qualified by feeling) which explains why the mechanism was selected and not why anything is left over once the selection story is complete. §7 supplies that half from projective geometry instead, and adds no property to do it.

S3.15 The relativistic theory of consciousness: frames, equivalence, and the individuation gap (§§7, 9)

§7 cites Lahav and Neemeh (2022) as an independent derivation of the Origin Thesis's central asymmetry, and §9 cites the same paper as the most permissive verdict on artificial consciousness now in print. Both citations are to one argument, and the argument is strong enough that the two uses have to be separated carefully.

The diagnosis comes first, and IFT endorses it. Naturalistic dualism and illusionism are presented as opposite responses to a shared and unexamined assumption: that phenomenal consciousness is an absolute property, so that a system either has it or does not, full stop. Relativity in physics denies exactly that form for a large class of properties. Velocity is not a

property a body has; it is a property a body has with respect to a frame, and two observers who disagree about who is moving are both measuring correctly. Lahav and Neemeh propose that phenomenal consciousness has the same logical form: a system has it with respect to a cognitive frame of reference, where a cognitive frame is fixed by the dynamics and representations of a cognitive system.

The positive argument is an equivalence principle. Let Alice be a conscious human and ALICE an artificial system with the same representations, memories, dynamics, and behavioral responses, stipulated to be a zombie. Because the two agree on the outcome of every measurement either can perform (including their own phenomenal judgments) the relativity principle requires that the same physical laws hold in both frames. If phenomenal consciousness is physical in the broad sense, and Alice measures a quale, then the transformation from Alice's frame to ALICE's must return a quale in ALICE's frame as well. The conclusion is that zombies are not nomologically possible, and that in the frame that generates them, phenomenal judgment-representations and qualia are identical rather than merely correlated.

The account then explains apparent privacy without invoking privacy. When Bob attempts to measure Alice's quale directly, his perceptual, cognitive, and linguistic subsystems return the empty set, because Alice's representations are not states of Bob's state space and have no causal power in his dynamics. He must fall back on his sensation subsystem, which registers only physical properties, so he obtains a brain pattern rather than an experience. Qualia are not private; they are frame-relative, and the transformation between frames is nomologically possible while remaining technologically unavailable.

The convergence with §7 is genuine and was arrived at independently. Both accounts locate the residue in a structural asymmetry rather than in an extra property, and both hold that the asymmetry is a consequence of how a format is constituted rather than a defect in introspection. Where Merker's projective formulation says that the origin of a projection is not an element of the projected field, the relativistic formulation says that a format is directly measurable only from within the frame whose dynamics it drives. These are the same observation in two vocabularies, and their independence is evidence that the observation is not an artifact of either.

The divergence is over what follows. Lahav and Neemeh conclude that the hard problem dissolves once relativity is accepted, since nothing further needs explaining. IFT holds that frame-relativity generates the residue rather than dispelling it. Granting that a system's format has causal power only within its own frame, that system has no representational resource for the frame's constitution, and the impression of something present and unspecifiable is the predicted consequence of the very relativity the theory establishes. The two accounts therefore agree on the mechanism and disagree on whether explaining why an intuition must arise counts as removing it. §7 marks this as the bet the Origin Thesis makes.

The disagreement that matters for §9 is constitutive rather than interpretive, and it is where IFT must actually defeat the argument. Three points.

First, the equivalence principle is applied to a pair whose individuation is assumed. Alice and ALICE are stipulated to agree on every measurement, and the relativity principle is then invoked to conclude that they are governed by the same laws. But whether an externally powered, checkpointable, replaceable process is an individual of the same kind as an organism is precisely what §9 disputes, and no measurement-agreement stipulation settles it, because the facts at issue (who acquires the energy, whose failure a fault is, whether a checkpoint restores the same token) are not facts about the measurements the two systems can perform on their own representations. Agreement on phenomenal judgments is agreement about the format and its consumers. It is silent on the index.

Second, the theory closes the consumer set by fiat, and closes it at a designated subsystem. The functional-consciousness subsystem C is defined as the module that receives percepts and produces phenomenal judgments, and phenomenal consciousness is then said to be present wherever C directly uses those representations in its dynamics. This is the higher-order move with a different label: a distinguished consumer is nominated, and nothing in the theory explains why that consumer rather than another well-placed module is the subject. §4's cardinality argument applies without modification, and here it has maximum force, because the verdict transfers automatically to any architecture in which some module consumes representations that carry content about phenomenology. DreamerV3 routes one latent to heterogeneous heads; BriSe AI builds five levels of self-model; a language model produces phenomenal judgments in quantity. On the relativistic criterion these are candidates. On IFT they are services.

Third, and most interesting, the theory itself supplies the counterexample. Having defined a cognitive frame by dynamics and representations, Lahav and Neemeh find that this is insufficient and must add spatial position. Alice and ALICE, sharing dynamics exactly, occupy the same cognitive frame yet not the same first-person frame, because each must use its sensation subsystem to measure the other and therefore obtains physical patterns rather than experiences. They name this the broken-symmetry scenario and treat it as a technical correction. It is not. It concedes that computational identity does not individuate the subject and that a further, token-level fact is required, which is §9's claim about computational invariance, reached from the opposite direction. IFT's objection is then narrow and specific: position is the wrong token-level fact. Two systems can differ in position while both remain externally maintained services whose cessation is a loss to an operator, and nothing about occupying a location makes a running process one whose failure is its own. Position individuates the process; it does not index it.

What would change the verdict is worth stating, since the argument is formally serious. If a transformation between cognitive frames were actually performed (if a system were reorganized to have the dynamics and representations of another and reported the other's

experiences) the relativistic framework would have delivered a prediction no competitor can match, and the stipulative character of the Alice/ALICE pair would cease to matter. Lahav and Neemeh are explicit that this is nomologically possible and technologically unavailable, and IFT does not dispute the possibility claim. The two theories can be separated before that becomes feasible, because they disagree now about ordinary artifacts: the relativistic criterion assigns qualia to any system with a phenomenal-judgment consumer, and IFT withholds them from every externally maintained one. F4 is that comparison, and the theories make opposite predictions about System A (S6.4).

S3.16 Phenomenal concepts and the Origin Thesis (§7)

The Phenomenal Concept Strategy (PCS) is the closest philosophical precursor to the Origin Thesis and the most important omission from the compressed discussion in §7. Balog (2012) accepts the epistemic and explanatory gap rather than explaining it away: physical and functional information does not tell a subject what a state is like. Her claim is that the gap concerns the concepts through which experience is grasped, not an extra nonphysical property. If that explanatory scheme is coherent, the gap cannot by itself carry the anti-physicalist conclusion usually loaded onto it.

On the constitutional account Balog defends, canonical first-person, present-tense phenomenal concepts are partly constituted by the token experiences to which they refer. The experience serves as its own mode of presentation. PAIN therefore differs from WATER not because pain is nonphysical, but because the concept does not present its referent by a causal, functional, or descriptive route. An instance of the referent is physically present in the concept token, which explains acquaintance, directness, and why a physical description of the neural state does not substitute for what it is like to token the concept. The same directness makes zombie scenarios conceivable without making them metaphysically possible.

This is a real advance over both crude reduction and crude illusionism. PCS explains why the explanatory gap persists in a physical world and why learning more neuroscience need not make the gap psychologically disappear. It also recognizes a limit IFT should accept: explaining why no reductive bridge is available is not the same as producing that bridge. Balog therefore presents PCS as an explanation of why a perspicuous reduction cannot be had, rather than the missing reduction itself.

Two naturalistic objections must be answered before this can count as an explanation. Churchland (1996) rejects the move from present inconceivability to metaphysical impossibility and argues for “reductive contact” between psychological and neurobiological levels rather than a demand for one transparent bottom-up deduction. Dennett (2001) predicts the extinction of the Zombic Hunch (the conviction that any mechanistic account must leave an ineffable residue) and his topic-of-consciousness question adds a constraint new physics does not meet: even sentient microphysics would not explain how myopic parts jointly represent Braque, poodles, or Kilimanjaro. IFT accepts both constraints. T4 is not an

argument from bafflement, and it does not exempt consciousness from multilevel neuroscience. It proposes a positive structural reason why one asymmetry remains inside a completed mechanism: the format can represent its contents, uses, and effects but cannot present its own origin as an element of the field it organizes. If mechanistic explanation removes that asymmetry, T4 is false; if it leaves only psychological reluctance, Dennett is right and T4 adds nothing.

The divergence is the order of explanation. PCS begins with a token experience q and uses its presence in the phenomenal concept Cq to explain the subject's epistemic position. Balog is explicit that a merely neurophysiological or architectural description of the concept's constitution cannot explain acquaintance unless the phenomenal presence of q is already granted. The strategy therefore addresses the gap between experience and description after the experience has been admitted. It does not identify the bearer of q, distinguish a token subject from a service, or explain why the medium that carries q is transparent to its users.

IFT places the asymmetry upstream. A projective format delivers content to its consumers while withholding its own constitution, and its origin cannot appear as an item in the field it organizes. A phenomenal concept formed within that format can therefore use an occurrent state as its mode of presentation without providing a description of the format that makes the presentation possible. The directness PCS treats as a special feature of phenomenal concepts is, on this account, a downstream consequence of format transparency and indexing. PCS explains why the gap survives once experience is given; the Origin Thesis predicts why an indexed format that models its own modeling generates the gap in the first place.

The accounts also make different claims about necessity. IFT does not require sophisticated phenomenal concepts for consciousness: a nonlinguistic animal may have an indexed format without the conceptual resources to formulate the gap. Conversely, a language model can deploy the vocabulary and inferential roles of phenomenal concepts without any token experience partly constituting them. Avowal therefore remains a bad indicator. What T4 predicts is conditional: once an indexed system acquires metacognitive access to the transparency of its own format, it should produce the sense of a present character that its physical self-description omits.

Balog's final symmetry is a constraint on how strongly that conclusion may be stated. She argues that PCS and anti-physicalism reach an a priori stalemate: each can block the other only by assuming its own contested starting point, and no neutral principle decides between them. IFT does not claim to win that ontology war by definition. T4 is a structural physicalist explanation whose merit lies in unifying the hard-problem intuition with the same architecture that T1–T3 use to explain format, bearer, and geometry. The future test in S6.5 is correspondingly not a proof of physicalism. It is the first empirical risk the account can take: an indexed artifact trained without phenomenal discourse should still generate the residue if the Origin Thesis is right. A robust absence would not establish dualism, but it would remove IFT's explanation of the gap.

Metzinger's (2003) self-model theory is the other precursor §7 names, and the relation is closer than to PCS. Phenomenal transparency, on his account, is the unavailability to introspection of the earlier processing stages of a representation, so that the representation is lived through as the world rather than attended to as a model; the phenomenal self-model is transparent in this sense, and its transparency explains why a system that is no one experiences itself as someone. T4 adopts transparency as the mechanism of the residue and rejects the conclusion drawn from the self-model, for the reason §4 gives against every self-model route (S3.10): the self-model is content, and the bearer is not. What transparency conceals is therefore not a fiction standing where a self should be but the constitution of the format whose contents the regime recruits. The two accounts agree that nothing extra must be added to the physical story and disagree about whether the subject survives the subtraction of the self-model; IFT's answer is that it does, because it was never the self-model.

S3.17 Universality, intrinsicness, and the Specificity Problem (§4)

§4 compresses a meta-theoretic argument that deserves its own statement, because it converts IFT's central negative claim from a complaint about particular theories into a general constraint on what a theory of consciousness has to be.

Kanai and Fujisawa (2024) propose universality as a desideratum alongside falsifiability. A theory is universal if it can determine, for any fully described dynamical system (biological brain, hurricane, or computer) whether that system is conscious, and if so which part of it is the conscious entity as opposed to its environment. Two clauses do the work. The first is scope: the theory must apply to any system whose dynamics are specified at a grain at least as fine as the level the theory concerns, so that the relevant features are detectable if present. The second, and the one IFT builds on, is intrinsicness: the determinant of consciousness must be a property the system possesses, not one assigned by an external observer's interpretation. They are careful to note that this is not a covert preference for computationalism. A theory holding that consciousness requires biological neurons can be universal, provided it states criteria by which an arbitrary physical system can be judged to contain neurons.

The criterion generalizes and tightens the fourth of Doerig et al.'s (2021) hard criteria, which requires a theory to cope with the multiple realization of functions across diverse entities. Doerig and colleagues count global workspace theory, predictive processing, and adaptive resonance theory as satisfying that criterion; Kanai and Fujisawa argue that none of them satisfies the stronger one, because coping with multiple realization in familiar cases is not the same as being able to decide arbitrary ones.

Their verdicts on the major functionalist theories are the verdicts §4 reaches by another route. Global workspace theory offers no mathematical statement of what constitutes a workspace, a broadcast, or an ignition in an arbitrary system, so it remains applicable only where a human neuroscientist recognizes the pattern. Higher-order theory offers none for

meta-representation, and the available options are both bad: a liberal definition on which a meta-representation is any transformation y = f(x) makes every layer of a neural network a meta-representation of its predecessor and slides into panpsychism, while uncertainty estimation as the criterion makes a softmax over an output layer sufficient, rendering most computer-vision networks conscious. This is the cardinality problem viewed from above. Where §4 says that "available to" has an unspecified second argument, Kanai and Fujisawa say the theory cannot decide arbitrary cases; these are one defect described at two levels of abstraction.

Their remedy and IFT's diverge, and the divergence is the substantive contribution. They propose that functionalist theories be made universal by rigorous mathematization of their constituent concepts. IFT denies that this suffices, and the denial is not a quibble. Suppose "global workspace" were given a fully general mathematical definition: a partition of a dynamical system into modules, a broadcast relation over them, and a threshold. Every one of those is a choice made by the analyst, and the theory would then decide arbitrary cases by applying the analyst's choices to them uniformly, which is determinacy, not intrinsicness. The same objection applies to the theory Kanai and Fujisawa award universality to. IIT does decide arbitrary cases, but it decides them by an exclusion axiom that selects one complex among nested candidates, and §4, §5, and S3.2 argue that the axiom is a chooser in the clothing of a postulate. Determinacy is cheap; intrinsicness is not.

Exactly one candidate determinant is intrinsic in their sense, and the reason it is intrinsic is that there is no vantage from which an observer could assign it. Which running token's failure is its own is not a fact about how a describer carves the system, because the carving is performed by the process itself in the course of continuing to exist. This is the sense in which enaction is not a biological preference but the only known non-stipulated cut. IFT is therefore universal on their criterion while being restrictive on the verdicts a permissive functionalism expects: it applies in principle to any physical system, and withholds consciousness from most of the systems such a functionalism would grant it.

Two further points of contact are worth recording. First, Kanai and Fujisawa scope their own criterion. Not every theory needs universality, and a non-universal theory can be perfectly serviceable in applied settings, predicting whether a patient under anesthesia can feel pain, for instance. Universality becomes obligatory when a theory aims at the hard problem, and at the harder problem of attributing consciousness to non-biological systems. IFT aims at both and so accepts the obligation, which is also why §12 is written as it is: a theory claiming universal scope owes falsification conditions in every domain it claims to cover, which is why F1 through F3 are biological and F4 is not.

Second, the criterion supplies a clean answer to the Specificity Problem, which Kanai and Fujisawa adopt from the philosophical literature on non-human minds and treat as the principal obstacle to theory-heavy attributions of consciousness beyond the human case. The problem is that theories keyed to high-level cognitive functions (working memory, attention,

metacognition) are keyed to functions admitting many realizations and many degrees, so extrapolating them to a novel system requires judgment calls the theory does not license. IFT is not exposed to that problem, because its criterion is not a cognitive function admitting of degrees. A running process either is or is not the thing whose failure the policy is regulating. This is not an advantage in ease of measurement, which §11 concedes is severe, but it is an advantage in kind: IFT's difficulty is epistemic, whereas the Specificity Problem is a defect in the criterion itself.

Start with the intrinsicality postulate, the strongest available rival stated at full strength. §4 argues against the current formulation of IIT rather than its 2012 predecessor for a plain reason: IIT 4.0 is the only framework in the literature that has turned the question §4 poses into an explicit postulate. Its methodological guidelines state that when accounting for experience in physical terms, existence must be evaluated from the intrinsic perspective of an entity (what exists for the entity itself) and not from the perspective of an external observer, and that from the intrinsic perspective the quality and quantity of existence must be observer-independent and cannot be arbitrary (Albantakis et al., 2023). Two ontological principles implement the demand. The principle of maximal existence holds that, when it comes to a requirement for existence, what exists is what exists the most; it is used to select the cause–effect state a system specifies, the set of units constituting a complex, the subsets composing its distinctions and relations, and the grain of the units themselves. The principle of minimal existence holds that nothing exists more than the least it exists; it is used to evaluate irreducibility over the partition across which the system is least irreducible, and to take the minimum of cause-side and effect-side irreducibility. A theory that states these principles is not vulnerable to the complaint that it has left the bearer tacit. It has an answer, and the answer is a maximization over intrinsic cause–effect power.

The postulate does not reach the argument's starting partition. The objection in §4 is not that the mathematics is wrong but that the intrinsic perspective is enforced from outside. The analysis begins from a substrate in a state with a transition probability function over its units; candidate systems are subsets of that substrate; and the remaining units are treated as background conditions and causally marginalized, conditional on the current state of the universe. Four things must be supplied before any postulate can be applied: which units the universe contains, at what spatial grain, over what update interval, and which of them count as background. The framework is explicit about the last point in practice, stating that a causal model will be neither complete nor atomic but will capture the features relevant to what the analyst is trying to explain and predict; and it is explicit about the cost, noting that a full analysis can be performed only on idealized systems of a few units, because the number of candidate systems, partitions, candidate distinctions, relations, and grains produces nested combinatorial explosions requiring many rounds of maximization. An intrinsic perspective specified relative to a stipulated boundary between system and background is a conditional intrinsic perspective. This is the same defect §4 identifies in availability accounts, arriving by

a different route: where a workspace theory leaves the consumer set open, an integration theory closes it by a rule applied to a partition the analyst chose.

Maximization is not individuation. The principle of maximal existence is a tie-break, and the framework says as much in its own terms: among candidate complexes overlapping the same substrate, the one that exists is the one laying the greatest claim to existence as one entity, and overlapping sets with lower system integrated information are excluded from existence for the same reason that gave the winner its status. That is a coherent decision procedure, and it presupposes the candidate set. Nothing in it explains why those units, at that grain, with those background conditions, are the alternatives among which existence is being awarded. IFT's alternative is not a better maximization. It is the observation that exactly one boundary is not supplied from outside: the one a process produces and must keep producing in order to continue (S3.2). The consumer set therefore closes at the regime whose competing demands resolve into one enacted commitment.

The count is an output, not a finding. The consequence §4 emphasizes is arithmetical. Because the search for maximal substrates is recursive (identify the substrate with maximum system integrated information, remove its units, search the remainder, repeat) a universe in a state condenses into a disjoint and exhaustive set of complexes: a first complex, a second, and so on. The number of subjects a system contains is therefore whatever the maximization returns over the description it was given. Two published examples make the scale concrete. In a six-unit network organized around a bottleneck unit, the analysis returns one two-unit complex together with four single-unit complexes, which the authors call monads: four subjects consisting of one unit each, in a network of six (Albantakis et al., 2023). In the brain, two or more non-overlapping complexes are held to coexist as discrete physical substrates of consciousness, each with its own definite borders and its own maximum of integrated information, with a major complex specifying the day-to-day stream and one or more minor complexes possibly supporting sophisticated performance that the major complex registers as unconscious; splitting of the major complex is predicted after callosotomy, canvassed for sustained dual-task conditions such as driving a familiar road while holding a conversation, and canvassed again for conversion and dissociative disorders (Tononi et al., 2016).

The divergent prediction is empirically testable. IFT's count differs, and not as a matter of taste. A callosotomized patient has one body, one internal milieu, one set of viability conditions, and one next act; the processes that maintain that active token and resolve its competing demands into one policy are not duplicated by the section of a commissure. IFT therefore predicts one subject with divided content and divided access, exactly the pattern §2's dispositional reading of the access condition permits, since a state may belong to the standing field while no consumer on a given side samples it. The empirical situation currently favors that reading. Across a wide range of tasks, patients with complete and radiologically confirmed transection of the corpus callosum show full awareness of stimulus presence and above-chance report of location, orientation, and identity throughout the visual field,

irrespective of whether the response is made with the left hand, the right hand, or verbally, while remaining unable to compare stimuli across half-fields: divided perception, one responding agent (Pinto et al., 2017). The finding is contested and IFT does not need it settled. What matters for §4 is that the two theories are committed to opposite answers about the same patients, that each answer follows from its individuation rule rather than from an added assumption, and that a verified case of two independently reporting agents with separately manipulable viability conditions would count against IFT in the way S6.2 specifies for the localization claim.

A second natural experiment runs in the other direction, and §12 pairs it with callosotomy. Craniopagus twins joined at the head and sharing, by report, a thalamic bridge are described as sensing stimuli delivered to the other twin, including taste and touch, and as able to see through the other's eyes, while remaining two persons with separate preferences, separate control of their own limbs, and separate viability (Dominus, 2011). A peer-reviewed neurological account of the case has not been published, and IFT does not rest on its details; what matters is the form of the case, content crossing between two regimes. IFT counts two subjects with partly shared fields, because there are two bodies with two sets of viability conditions and two next acts. A theory that counts by maximal integration must either find the bridge insufficient and agree, or find it sufficient and count one. Together with callosotomy the cases form a two-by-two design, content shared or divided against regimes one or two, in which IFT's count follows the regime in every cell and the integration count follows the connectivity. Complete callosotomy has now been shown to disrupt interhemispheric network interaction in a way that partial section, even to a centimeter of splenium, does not (Santander et al., 2025), which sharpens the first cell without settling it; the behavioral question remains whether the disconnected hemispheres are two reporting agents with separately manipulable stakes, and S6.6 states the criteria.

What would change the verdict is exact. If a formulation of intrinsicality derived its units, grain, update interval, and background conditions rather than receiving them from an analyst, the meta-theoretic objection in §4 would lapse, and the disagreement with IFT would become first-order. IFT does not answer by maximizing a rival scalar. Index evidence first fixes the bearer; interventions then identify the format's causal grain: its constitutive units are those whose perturbation produces content-specific reorganization across discrimination, valuation, memory, and policy, and its update grain is the interval over which that reorganization propagates and settles. Physically large differences below that grain may be phenomenally silent. This rule is operational rather than ontological (the persistence regime individuates the subject, intervention maps the format within it) and it is fixed before the phenomenal verdict. Kanai and Fujisawa's (2024) universality requirement is the right demand; §4 adds that no amount of mathematization can meet it while the arguments of the criterion remain chosen (S3.19).

S3.18 Engineering requirements converge on the five factors (§2)

§2 claims that the five empirically dissociable factors are the parts of one control problem rather than an accidental list. The claim is strengthened by a convergence from a literature with no stake in it.

Dehaene et al. (2017) derive the same commitment pressure from C1. Specialized systems can maintain multiple probabilistic interpretations, but an organism or autonomous agent must pool evidence, select one course, sustain it across delay, and coordinate internal and external processes around it. This is the cleanest functional precursor to policy closure. Their account stops at the architecture: the same C1 design can be installed in a car or service whose fuel, repair, and continuation remain somebody else's problem. IFT adds the missing possessive. A commitment closes constitutively only when it is the next act of the active token whose continued organization the policy regulates.

Rosenbloom et al. (2026) make the convergence architectural. Mapping GNWT, IIT, RPT, and predictive processing/neurorepresentationalism onto the Common Model of Cognition, they recover recurrent local modules plus a cognitive cycle operating on a global working-memory hub with complex, possibly multimodal states. GNWT maps to broadcast through working memory; RPT to local and global feedback; predictive processing to recurrent hierarchies and high-level multimodal representations; IIT to the integrated structure of a current state. The point is not that the mapping proves any theory. It is that four purportedly constitutive theories can be expressed as overlays on one consensus architecture of humanlike cognition. The Common Model is explicitly abstract and incomplete, and emotion and metacognition remain proposed extensions. It therefore describes a package of format and consumer functions while leaving bearer, valence, and individuation untouched.

Prokhorenko et al. (2026) provide the implementation-level control. Their temporal predictive-coding world model learns compact latent states under partial observability, uses recurrent prediction errors to stabilize temporal memory, distinguishes hidden grid positions that share the same observation, generates internal rollouts for candidate actions, and supports reward-guided policy selection. Performance and latent-state generalization improve as the agent learns. This is the machinery most often bundled under world modeling, temporal thickness, recurrence, memory, and planning, realized without any appeal to subjectivity. The task, reward objective, training procedure, hardware, and continuation of the running instance remain externally specified. TPC therefore strengthens the orthogonality claim: stable predictive internal states are evidence for a format and its consumers, not for an index.

Wray et al. (2026) ask what a general agent needs in order to behave sensibly in situations outside its design scope, situations qualitatively rather than merely statistically unfamiliar, for which no policy was trained. They distinguish three cases: out-of-distribution situations, where the features are apt but their combination is novel; out-of-designed-feature-scope situations, where observations contain data salient to a good response but the agent has no

features for them; and out-of-observation situations, where the agent's embodiment cannot sense what should govern its action. Their requirements are derived from the problem rather than from any theory of consciousness, and they reproduce the constraints §2 says generate the factorization. The agent must survive in fail-hard environments, where a single failure may destroy its embodiment, so adaptation by repeated exposure is unavailable. It must act under irreversible dynamics, since resource consumption and physical change make the safely explorable state spaces assumed by many algorithms unavailable. It must persevere zero-shot, responding on a first and possibly only encounter. It must mitigate partial observability, scale to long-horizon tasks whose multiple objectives may conflict, and adapt at whatever timescale the moment allows, reacting fast when speed is what the situation demands and deliberating when it is not.

Their proposed solution is more interesting still. Having reviewed and rejected broad prior knowledge, extended deliberation, predefined fallbacks, return-to-familiarity, slow learning, and ensembles of these, they turn to human appraisal processes: fast, domain-general commentary on relevance, novelty, urgency, goal-conduciveness, coping potential or control, and normative significance, generated alongside task reasoning and used by metareasoning to select which decision strategy the situation warrants. That is, in a different vocabulary, the architecture §2 describes. A global state that permits modeling at all; a structured model; a shared format from which processes recruit; a commitment that must close, because one embodiment has one next act; and an annotation of how far the model may be trusted—with appraisal doing the work §4 assigns to valence, converting a situation into a demand that is ranked against competing demands rather than summed with them.

What the convergence does not supply is the index, and the paper is unusually clear about where the gap sits, though it does not describe it as a gap. The requirements are desiderata the designers wrote down. Survival appears first because the authors put it there, not because anything about the agent makes its cessation its own failure; an agent that fails to survive has cost its operator an asset. Coping potential and control, the appraisal dimensions carrying the most weight in their analysis, are estimates a system computes about its own prospects, and a checkpointable service can compute them accurately while nothing whatever follows for it from a low value. This is the point §4 makes against MetaMo's designated safe region (S3.3) and against Zeng et al.'s hierarchical self-models (S3.10), arriving from a third direction: an appraisal of stakes is a representation, and a representation of stakes is not stakes.

The prediction that follows is testable in the ordinary engineering sense and forms a component of F4 (S6.4). An appraisal-equipped agent whose maintenance is external should show the appraisal dimensions operating as annotations that modulate strategy selection, and nothing further; the same architecture in a constitutively mortal device should show appraisal reorganizing attention, learning, memory, and action across domains, because the appraised quantity is then not an estimate about a system but a condition of the estimator. Cross-domain counterfactual depth is the discriminator here, as it is throughout §11.

S3.19 The interpretation problem: agency and resources as two-place relations (§4)

§4 argues that informational, statistical, and causal criteria cannot individuate a subject because each must be handed the cut it scores. Formal agency theory sharpens the result: the partition is chosen, and the hidden state, objective, action, and resource boundary can be chosen with it.

Biehl and Virgo (2023) set out to say precisely when a physical system may be regarded as an agent with beliefs and goals. Their target system is a stochastic Moore machine: a set of internal states, an input space, a machine kernel taking an input and a current state to a distribution over next states, an output space, and a deterministic expose function from states to outputs. To this they add a consistent Bayesian influenced filtering interpretation, consisting of a hidden state space, an action space, an interpretation map from machine states to distributions over hidden states, an action function from machine states to actions, and a model kernel giving the joint dynamics of next hidden state and next input given a hidden state and an action. The interpretation is consistent when the belief assigned to the next machine state is the Bayesian posterior of the belief assigned to the current one, given the input received in between. A system counts as solving a partially observable Markov decision process when it carries such an interpretation with respect to the POMDP's transition and sensor kernel and its expose function returns, for every machine state, the action the optimal policy assigns to the belief that state parameterizes. An agent, on their proposal, is the pair of a system and an interpretation of it as a POMDP solution.

The construction is careful, explicitly Dennettian in inspiration, and offered by its authors as an example of the kind of answer the question requires rather than as a complete theory. It is also, read against §4, a demonstration. Three of its features are the ones that matter here, and none is incidental to the formalism.

First, the hidden states and the rewards are not assumed to have any physical realization. They are part of the interpretation and, in the authors' own words, need only exist in the mathematical sense. The hidden state is assumed by the agent to exist and need not correspond to any state of the true external world. A goal, on this construal, is an element of a description an analyst writes down, and the same running machine may be a rational solver of one POMDP under one interpretation and of another under a second.

Second, the machine never observes the rewards by which its rationality is assessed. Optimality is computed by the interpreter, over a reward function the interpreter supplies, using an optimal value function the interpreter derives. Nothing in the machine registers whether reward was obtained. This is the formal counterpart of §4's claim that assigning a scalar objective is not conferring stakes, and it is worth noting that the point is not a criticism of Biehl and Virgo but a feature they build in deliberately: the interpretation is intended to be about beliefs and goals ascribed to a system, not about anything the system undergoes.

Third, and most striking, consistency depends on the machine alone and not on any property of its environment. The authors emphasize that a consistent interpretation requires the agent's beliefs and goals to be consistent rather than correct, so an agent may hold a model that does not correspond to its actual world and act suboptimally with respect to that world while remaining an agent by the definition. They extend the observation to action: a machine with no outputs at all can carry a consistent interpretation on which its internal state influences the hidden state dynamics, even though it can have no influence on the actual environment. Such actions, as they put it, remain confined to the interpretation. A further feature completes the picture: interpretation maps need not be injective, so coarse-graining of the internal state is built into the notion of interpretation itself, which is the partition problem of S3.2 arriving from the same direction.

The consequence for this paper is direct. Availability, closure, integration, and now agency itself are all relations whose second term is furnished by whoever is doing the describing. The partition problem showed that the cut is supplied; the interpretation problem shows that the objective, the hidden variables, and even the actions can be supplied. It is not that these formalisms are defective. It is that they are formalisms of ascription, and ascription is what a constitutive theory of subjecthood must not rest on. Note also the provenance: Biehl is a co-author of the information closure theory that §4 criticizes for being unable to decide whether there is one consciousness, two, or three (Chang et al., 2020). The same research program has produced both the closure criterion and the formal reason why criteria of that shape cannot individuate.

Fuller et al. (2026) prove the same underdetermination inside a formal intelligence measure. They define resource helpfulness in an agent-focused form by bundling π with a reward-free resource R to form a composite agent πR, and in an environment-focused form by wrapping every environment μ with the same resource to form μR. Their Theorem 1 shows that the two constructions yield identical expected reward and therefore identical resource-relativized Legg–Hutter intelligence. Figure 2 makes the point visually: the same R lies inside the dashed boundary of the composite agent under one description and inside the dashed boundary of the composite environment under the other, while the interaction history is unchanged. The theorem is not a theory of consciousness, and its authors do not present it as one. Its relevance is sharper: performance cannot decide whether a memory, oracle, tool, retrieval store, or auxiliary computation belongs to the agent or merely surrounds it. A metric of goal achievement can therefore be exact while the bearer remains description-relative. The resource boundary joins the cut, objective, hidden state, and action as another term an interpreter can move without changing the competence being measured.

The same as-if structure has now been pushed down to physics. In the gauge-theoretic recasting of the free energy principle surveyed by Safron et al. (2023), mechanical theories are specified for systems that look as if they are estimating posterior distributions over the causes of their sensory states. That formulation is honest and, on IFT, exactly right about its own

scope: it supplies a description under which a system can be read as inferring, and a description under which a system can be read as inferring is not a fact about what any system is undergoing.

What IFT adds is the invariant. The indexing relation is not interpretation-relative, because it is not a relation between a system and a description but between a running token's states and the conditions of that token's own continuation. Those conditions are not selected by the analyst; they are enacted, and an analyst who changes them changes or destroys the individual rather than re-describing it. This is also the precise content of the universality demand discussed in S3.17: a determinant the analyst cannot choose is a determinant invariant across admissible interpretations, and Biehl and Virgo make vivid how large the space of admissible interpretations is for every criterion stated in terms of beliefs, goals, or availability. The methodological consequence is recorded in F4 (S6.4), where index status must be established from constitutive evidence about power, repair, checkpointing, and replacement, before any behavioral or architectural comparison is run.

S3.20 Levels: multi-scale individuation and developmental scaffolding (§4)

The Indexing Thesis fixes a level, and a theory that fixes a level owes an account of the two hardest cases: systems whose parts are themselves agents, and systems whose regulation is initially performed by somebody else. §4 answers the first with the bearer/beneficiary distinction and Ginsburg and Jablonka's level clause. Two active research programs press harder than that clause alone can bear, and both are worth stating in full because IFT's answer in each case is substantive rather than protective.

The first is Levin's program on multi-scale competency, which asks IFT's question in IFT's terms. Pio-Lopez et al. (2023) put it directly: what evolutionary dynamics enable individual cells to integrate their activities so that a novel, higher-level intelligence emerges with goals and competencies belonging to it and not to its parts? Their answer runs through homeostasis. Using the French flag problem as a model of establishing and maintaining a body-wide positional axis, they show in simulation that emergent morphogenetic agents can combine local and global signaling to reach target morphologies, recover from perturbation, hold long-term stability, and remodel suddenly long after stabilization; they then observe related phenomena in regenerating planaria. Fluctuating stress levels, rather than a designed controller, do much of the coordinating work. Safron et al. (2023) place this alongside the gauge-theoretic and symmetry-breaking literature in which morphogenetic fields, Kantian wholes, and free-energy minimization are treated as descriptions of how parts come to serve wholes.

IFT takes the program as an ally on mechanism and parts company on criterion. What Levin's group describes is how a new policy-integrated persistence regime gets assembled: the process by which sub-agents with their own homeostatic set-points come to be regulated by, and to regulate, a set-point that is not any of theirs. That is precisely the assembly IFT requires

and cannot itself supply. But scaling of goals is not by itself scaling of subjecthood, and the criterion does not move with the amount of cross-scale signaling. A tissue in which competent cells coordinate elaborately, and each cell's viability is nonetheless the only thing that can fail, has many indices and no organism-level one. The transition IFT names is the appearance of a failure mode that belongs to the higher level: a state of affairs in which the demands of the parts must be resolved into one commitment because there is now one thing whose continuation is at issue. Multi-scale competency is the road; the closure of competing demands into a single policy is the arrival. This also explains why IFT can accept, without embarrassment, that cells within an animal retain individual viability regulation: the organism-level index does not require the abolition of lower-level ones, only that there be a level at which boundary maintenance, modeling, valuation, and action selection close into one failure mode.

The second program runs the other way, and it concerns onset rather than scale. Santaguida and Bergamasco (2023) develop an active-inference account of attachment in which the generative model begins forming before birth. Homeostasis, on their account, is the first prior; the values to be held stable are largely learned during a period of co-embodiment in which fetal and maternal systems are homeostatically coupled, with the mother's evaluation of her circumstances reaching the fetus through hormone secretion and the fetus tuning its own set-points to information received from a shared environment. More complex models are then laid down around that core like the rings of a tree: allostasis, proximity-regulation strategies, cognitive and behavioral emotion regulation, embodied internal working models, and finally attachment style. They end by asking, as an open question, where the boundaries of the infant's Markov blanket lie when the infant does not yet distinguish herself from the mother.

IFT's answer is that scaffolded is not stipulated, and the distinction is worth three cases. Dependence on an external regulator does not defeat an index: an organism whose oxygen, warmth, and nutrients arrive from outside is not thereby a service, and neither is a fetus, provided its own operation is what maintains it as the thing it is. A genuinely undivided regime would have one index rather than two, and the interesting empirical question raised by the co-homeostasis literature is exactly when the fetal regime becomes separately closed rather than a component of a maternal one. A service relation is different from both, and the difference is not that others decide its fate, since a ventilated patient and a prey animal are both at others' mercy and both have everything to lose. What disqualifies a service is that it has nothing of its own to lose: its identity is defined at the level of an owner or orchestrator for whom cessation is a cost rather than an ending, so that stopping it and restoring it elsewhere ends nothing (§9). Only the third case is excluded by T2.

Two further consequences follow. First, the developmental material supplies the strongest available answer to the gradedness threat identified in S2.7. IFT holds that the index is not a matter of degree, and development appears to be a counterexample because everything about

an infant arrives by degrees. It is not one. The format and its consumers are graded throughout life; what is not graded is whether there is a regime whose failure would be its own. What development contributes is indeterminacy at a boundary rather than a continuum of partial indices, and indeterminacy at a boundary is what any constitutive transition exhibits. Second, the corroboration runs in IFT's favor on the order of construction. If consciousness were constituted by broadcast, monitoring, or self-modeling, one would expect those to be the developmental core, with homeostatic regulation as an early peripheral competence. The developmental literature reports the opposite order, with homeostatic regulation as the first prior and representational elaboration built outward from it—the order §4 predicts and Damasio's hierarchy describes.

S3.21 Two conditions, not two types: the framework IFT completes (§2, §6)

§2 adopts, and §6 partly resists, the most consequential recent reframing of Block's distinction. Because IFT's relation to it is close enough to be mistaken for agreement and sharp enough to generate a preregistered disagreement, the position deserves its own statement.

Mudrik et al. (2025) begin from the divide between sensory theories, which hold that the right kind of processing within sensory cortices suffices for consciousness, and cognitive theories, which hold that non-sensory machinery is required for sensory content to be experienced. They observe that recent evidence, including the no-report literature and the Cogitate collaboration, supports parts of both and refutes the strong form of each, and they argue that the impasse traces to Block's proposal that phenomenal and access consciousness are two types of consciousness. Their correction is that P and A are two necessary conditions for consciousness: a conscious state must have potentially phenomenal content, structured so as to provide distinctive qualities from a first-person perspective, and that content must be accessed, integrated with other systems so as to make at least a minimal difference to cognition or behavior. A state that meets only one condition is a form of unconscious processing. They then draw the consequences: most existing theories are incomplete because each supplies a mechanism for one condition and leaves the other blank; the front-versus-back dispute should be settled with an "and" rather than an "either"; early and late correlates mark stages of one process, as trichromatic and opponent-process accounts marked stages of color vision; consciousness has at least the functions that access affords; and conscious perception may overflow report but not access.

Almost all of this IFT endorses, and the endorsement is not grudging. The claim that theories should be arbitrated within the content and access domains and unified across them is the reading §8 gives Cogitate. The insistence that access alone is not a kind of consciousness (that an artificial system with access to its internal states is not thereby conscious) is §4's conclusion reached from another direction. The refusal of the sensory sufficiency claim is the cerebellar test's conclusion without the cerebellum. And the demand for a theory that

explains both what makes content feel the way it does and how that content makes a difference is T3 plus T2.

The framework is nonetheless incomplete in a way its authors half-record. In their discussion of what access means, they note that access is a transitive verb implying something doing the accessing, disclaim any homuncular reading, and give the intended one: information encoded in certain subsystems is available to other subsystems that have their own relationship with behavior. They then state that exactly which subsystems these are remains a matter of contention which they do not aim to settle. That sentence is the cardinality problem of §4 written by the framework that most needs it solved. A necessary condition stated as a relation with an unfixed second term does not yet decide any disputed case: it cannot say whether cerebellar content is accessed by the relevant subsystems, whether an enteric plexus accesses its own states, whether a stored-program machine's registers access one another, or how many subjects a split brain or a colony contains. T2 is the missing term, and once it is supplied the two conditions become determinate: the accessing systems are the consumers standing in the policy-integrated persistence regime, and no others.

Dehaene (2014) supplies the precursor distinction but not IFT's verdict. His preconscious bodily signal is "accessible but not accessed," processed enough to guide posture yet unconscious until broadcast. That taxonomy establishes that standing availability and episodic recruitment are separable. IFT's claim is that, within an independently indexed field, the former is format membership rather than a waiting room.

One disagreement then remains, and it is exactly locatable. Mudrik et al. read the A condition as occurrent (content is conscious when an access mechanism has been applied to it) and infer that conscious perception cannot overflow access. IFT reads it as dispositional: format membership is standing recruitability by the policy regime, and episodic recruitment is a further event that may or may not occur. Three considerations favor the dispositional reading. The first is Cogitate itself. If access is occurrent, then the posterior structure that carried task-irrelevant orientation information throughout stimulus presentation, while prefrontal decoding sat near chance, was unconscious for the duration; but the same authors want posterior structure to be doing constitutive work in their "and" account of the front-back dispute, and it is hard to have both. The second is their own concession that access overflows report. Once a gap is opened between what is conscious and what is used for the paradigm case of use, the dispositional reading extends the same move one step rather than introducing a new kind of claim. The third is the cerebellum: an occurrent condition risks making the exclusion depend on which consumers happened to sample, whereas the dispositional condition excludes cerebellar content because it is not recruitable by the regime at all, which is the stronger and more stable result.

The disagreement is empirical and is registered as such. Protocol F1 (S6.1) now carries four preregistered accounts. Higher-order inflation predicts no effect of perturbing never-recruited posterior structure; IFT predicts a relation-specific effect even when an independent

post-cued assay shows the query service intact; the two-conditions account predicts an effect only where access subsequently follows; and Perera's (2026; S3.23) virtualization account predicts that any effect is mediated by impairment of the service that generates content on demand. The delay manipulation, structural-error pattern, and query-service assay separate the four predictions, and separate likelihoods are recorded for each.

A final point of contact deserves recording because it bears on T4. Mudrik et al. explicitly hold that both conditions are characteristics of mental states which, on a physicalist stance, are physical states, and that phenomenality can be studied with objective measures despite being subjective. IFT agrees, and the Origin Thesis explains the residue that survives the agreement: the reason phenomenality feels like something over and above the two conditions is not that a third condition is missing but that the origin of a projective format cannot appear inside the field it projects.

S3.22 Affective taxis: a formal model of valence without an owner (§4)

Sennesh and Ramstead (2026) offer the strongest formal challenge in the current literature to IFT's claim that scalar reward is the wrong kind of object. Their affective-taxis hypothesis traces valence to the evolutionary coupling of bodily movement and internal physiology. Interoceptive state sets the salience of attractant and repellent gradients; apparent valence is the sign and magnitude of the directional derivative of an allostatic energy function in the organism's current direction of movement. They formalize the system as a POMDP and propose energy-based implementations, extending literal taxis in physical space to an internal interoceptive landscape. This is not a variable merely named pain. It is a mechanistic proposal connecting physiology, action, and evaluation.

Three distinctions in the proposal matter. It separates valence from arousal; the present taxis landscape from the later associative critic that predicts its long-run course; and physical reward objects from the physiological condition that makes them attractive or repellent. IFT accepts all three. They undermine one-currency reinforcement stories and support §4's claim that value is context-sensitive, body-relative control rather than the quantity of an external good.

The disagreement is about constitution. A signed derivative remains a scalar, and the POMDP can be specified for a worm, a robot, or an observer's model of either. The paper's alignment aim is for artificial agents to represent human affect and its interoceptive grounds, which states the gap cleanly: representing an affective landscape is not having one. Unless the coordinates are conditions of failure for the same active token whose policy the gradient steers, the gradient is a control signal in a service. Biehl and Virgo's interpretation result reappears: state variables, rewards, and actions can be assigned under a model; no equation identifies the bearer.

IFT therefore treats affective taxis as a candidate biological vehicle and a stronger engineering test, not as an index. Give System A an allostatic POMDP whose internal variables are sensed,

predicted, and used to organize behavior while power and repair remain external; IFT predicts no subject. Give System B the same architecture, but make those variables conditions of failure for the non-checkpointed substrate and close them into one policy; IFT predicts the dissociation manifold. F3 supplies the direct challenge: if manipulating gradient geometry reverses hedonic sign while the physiological stake is held fixed, valence is topological and T2 fails.

S3.23 Virtualized subjectivity: a computational rival to standing geometry and an independent route to the Origin Thesis (§§6–7)

Perera (2026) begins from phenomenal functionalism: a science of subjectivity needs to explain why an agent takes itself to inhabit a rich, private phenomenal world, and sufficiently rich self-simulation is proposed as constitutive rather than merely report-generating. The account matters to IFT because it converges on two of IFT's results and exposes two exact disagreements.

The first convergence is the content–implementation distinction. Perera's virtual-matrix analogy shows how a sparse service can answer arbitrary queries over a dense abstract space without storing every element in a saturated data structure. Applied to experience, saccades, attentional shifts, and other subpersonal inferences produce content on demand while the personal-level agent consumes a seamless egocentric world. This is the strongest computational alternative to the Geometry Thesis identified here. It agrees that sparse machinery need not imply a sparse personal-level world. It denies that the full relational field must exist as standing substrate. On IFT, operational availability of a coordinate or relation is not enough: phenomenal geometry is the topology of the indexed format, and unqueried relations are constituents whose perturbation should alter later report. A virtual service predicts a later error only insofar as the perturbation impairs the mechanism that generates the queried content. F1 registers that as a separate alternative rather than folding it into higher-order inflation or the two-conditions account.

The second convergence is transparency. Perera locates apparent privacy at an abstraction boundary between subpersonal producer and personal-level consumer. The consumer receives world-as-seeming while the machinery that fixes that seeming is unavailable from its side of the interface; implementation therefore disappears and phenomenal properties seem intrinsic. This is the Origin Thesis in computer-science vocabulary. It supplies an architecture for why the medium cannot appear as content and why privacy is de facto rather than metaphysical. Perera also predicts that an artificial subject should find hard-problem and other-minds intuitions compelling, independently anticipating the T4 test of S6.5.

The disagreement is over what self-simulation constitutes. Perera closes with the claim that building a conscious machine means building one that thinks it is conscious—not by emitting a sentence, but by sustaining an egocentric self-simulation. IFT refuses the identity. A self-simulation is representational content plus metacognitive consumer relations; it can be

implemented in an externally powered, checkpointable, replaceable service. It may explain first-person reports, perceptual vantage, covert deliberation, and the intuition of privacy while leaving the cardinality question untouched: which running physical process is the one for whom the virtual world is present? The abstraction boundary identifies producer and consumer roles inside an architecture, not the boundary of the individual whose failure is its own.

Perera's proposed indicators (ownership of a first-personal vantage point, sustained reflective and covert behavior, and susceptibility to the non-functionality fallacy) are therefore serious tests of format and C2, not of the index. They belong in the dissociation manifold and are stronger than prompted avowal, but remain reproducible by a service whose identity is set by an orchestrator. The paper is consequently both ally and adversary: ally to the format/consumer distinction and T4 on transparency; adversary to T3 on whether density requires standing topology; and direct opponent to T2 on whether self-simulation can constitute a bearer. Each overlap is now a discriminating prediction.

S3.24 Nearest neighbors: enactivism, organizational closure, and the beast machine (§4)

§4 derives the bearer from self-production and policy closure, and three traditions have traveled part of that road. The relation to each is set out here so that the derivation's debts and its departures are both on the record.

The enactive tradition is the oldest neighbor and the one with the strongest claim on the first half of the derivation. Weber and Varela (2002) argued that an autopoietic system is the only kind of thing for which a Kantian natural purpose is more than a regulative idea, because its identity is the product of its own activity, and that sense-making, the appearance of significance in an environment, follows from that identity. Di Paolo (2005) showed that autopoiesis alone cannot ground norms, because a system that merely is or is not alive has no graded relation to its own viability limits, and added adaptivity: the capacity to regulate the system's own states with respect to those limits, so that some trajectories are better or worse for the system before they are fatal. Thompson (2007) made the resulting continuity between life and mind the thesis of a book. Barandiaran et al. (2009) then defined agency by three conditions, individuality (a self-produced boundary), interactional asymmetry (the system rather than the environment is the source of the coupling's modulation), and normativity (the modulation answers to norms the system's own existence generates), and Di Paolo (2009) named precariousness, the fact that the processes of a living system would run down without the system's own activity, as the condition that makes those norms more than metaphor. Froese and Ziemke (2009) drew the consequence for artificial systems: an artificial agent in the full sense would have to be constitutively autonomous and precarious, which no system built on an externally maintained computational platform is.

IFT accepts every item in that inventory as the account of what a bearer is, and §4's fourth and fifth premises are enactivist premises with one addition, policy closure, which fixes the level at which nested self-producing systems become one subject. The departure is on what follows. The enactive tradition treats sense-making as the minimal form of mind and reads consciousness as its elaboration, so that the difference between a bacterium and a primate is one of degree along a single axis. IFT denies this. A bacterium is a paradigm sense-maker and has no world: nothing in its organization is a relational structure recruitable, as structure, by the processes that close its policy, because there is no such structure to recruit. The gradient from bacterium to primate runs through the format, and the enactive literature has no account of the format, no account of why phenomenal structure has the character it has, and no engagement with the consumer-set problem that availability theories face. That is why §4 presents the derivation as half a theory: the enactivists supplied the bearer and stopped.

Organizational closure is the formal descendant of autopoiesis and the formal version of §4's fourth premise. Montévil and Mossio (2015) define a biological organization as a set of constraints that depend on one another for their maintenance, such that each constraint is produced by processes that other constraints in the set canalize and the set as a whole achieves closure; Moreno and Mossio (2015) build an account of biological autonomy on that notion, with regulation as the capacity of a closed organization to modulate its own constraints in response to perturbation. Neither takes a position on consciousness, and IFT does not need them to. What §4 borrows is the demonstration that a boundary can be a fact about a system rather than about a description, because the system's persistence depends on it: a Markov blanket can be drawn around anything, whereas a closure of constraints either obtains or the system is not there.

Seth's beast-machine account is the nearest neuroscientific neighbor, and the one most likely to be mistaken for IFT. Seth and Tsakiris (2018) ground selfhood in the predictive regulation of the physiological condition of the body: the deepest layer of the self is the brain's control-oriented inference about interoceptive states, and experiences of emotion, mood, and embodied selfhood are the contents of that inference. Seth (2021) extends the account to consciousness generally, and Seth (2025) to the restrictive verdict on digital artifacts (S2.1). IFT agrees with the direction of grounding: the subject comes from regulation, and the body that is regulated is the thing whose fate is at issue. It disagrees with the identification. Interoceptive inference is a service that a replaceable system can run, as the installed-valence architectures of S3.3 and S3.22 show, and the content of experience is mostly not about the body: the visual field, speech, music, and arithmetic are in the format without being interoceptive. On IFT regulation individuates the bearer and the format constitutes the content, and the beast-machine account has the first without separating it from the second. The disagreement surfaces in F4, where System A can run every interoceptive inference System B runs and is predicted to have no subject.

S4 The Cogitate result: additional detail (§8)

The Cogitate Consortium's preregistered adversarial collaboration (Cogitate Consortium et al., 2025) had proponents of integrated information theory and global neuronal workspace theory specify divergent predictions and pass/fail criteria in advance, across fMRI, MEG, and intracranial EEG in a common paradigm.

Posterior cortex sustained decodable stimulus category for the full stimulus duration in all three modalities. Prefrontal decoding was weak and confined to approximately 0.2–0.4 s after onset, and no electrode showed the onset-plus-offset ignition profile that the workspace prediction required. The orientation result is more diagnostic than the category result for the purposes of §8, because orientation was task-irrelevant throughout: it decoded reliably from posterior sites and sat at or near chance prefrontally, with Bayes factors favoring the null. A feature that no consumer was recruiting was nonetheless held, stably and decodably, in posterior cortex for as long as the stimulus was present—which is what §3 and §6 predict a format should do and what an access-identity theory does not require.

The preregistered phase-synchrony analyses failed the integrated information prediction with equal clarity: the sustained posterior gamma coupling supposed to specify conscious content did not appear. §8 argues that this was a prediction the theory should never have made, because its own flagship structural result locates content specification in standing connectivity rather than in dynamic coupling (Haun & Tononi, 2019), and the paradigm neither manipulated standing connectivity nor could have detected its contribution.

The Duhem–Quine point in §8 is worth stating in its strong form. Preregistration constrains the proponents who sign it; it does not constrain the space of alternative explanations. Whyte et al. (2026) review an active-inference model in which reporting is a mental action that gates second-level precision. Under no-report conditions, the model produces reduced prefrontal-like activity and no P3b-like response while maintaining a high second-level posterior. It did not predict the Cogitate experiment as such and therefore did not "win" the collaboration. It shows that the report-dependent half of the observed pattern was available from a framework not represented in the contest. Unless centrality weights and a wider candidate set are fixed in advance, an adversarial design cannot tell whether a live explanation was never at the table (Lakatos, 1970).

S4.1 The REM precedent: a correlate promoted to an identity (§8)

§8 argues that the Cogitate pattern is what the format–consumer distinction predicts, and closes by observing that the field has made the identification of a phenomenon with its correlate before. The episode is given here because it is the most fully documented case available and because its resolution has the same shape as the present account (Solms, 2021).

Aserinsky and Kleitman applied electroencephalography to sleep in the early 1950s and found, against expectation, that roughly every ninety minutes the brainwaves of a sleeping subject accelerate to near-waking frequencies while the subject remains asleep—paradoxical sleep, later renamed for the rapid eye movements that accompany it. With Dement they then tested the reasonable hypothesis that this state is the physiological basis of dreaming, and confirmed it: about 80 percent of awakenings from REM produced dream reports against fewer than 10 percent from non-REM sleep.

What followed is the substitution. Having established a correlation, the field treated it as an identity, and having an objective marker in hand, discarded the subjective side of the correlation as methodologically inconvenient—retrospective, single-witness, verbal, unreliable. Thereafter REM sleep was studied largely in animals that cannot report, and the neurophysiology proceeded on the assumption that explaining REM would explain dreaming. Jouvet localized REM generation to the pons, Hobson identified the pontine cell groups involved, and the inference drawn from the clockwork regularity of the cycle was that the motivating force for dreaming is physiological rather than psychological and that dreaming is motivationally neutral.

The identification was false, and the error was one of theoretical form rather than measurement. Dreaming and REM sleep are doubly dissociable. Patients with damage to the mesopontine tegmentum, the structure that generates REM, continue to report dreams; dreaming is abolished by damage in two entirely different places, the white matter of the ventromesial quadrant of the frontal lobes and the inferior parietal lobule, with REM sleep preserved. The two phenomena are correlated and are not the same thing. Corroboration ran through an independent literature: prefrontal leucotomy, which cuts the ventromesial frontal white matter, was documented to abolish dreaming, and the anatomy implicates the mesocortical–mesolimbic dopamine pathways coursing beneath the cortex there. Pharmacological evidence points the same way in both directions, since dopamine agonists increase the frequency, length, and intensity of dreams without commensurate effects on REM, while blocking the cholinergic system Hobson credited with generating dreams increases dreaming rather than reducing it.

The shape of the correction is what §8 draws on. The phenomenon was relocated from the mechanism that sets the global state (a cholinergic on/off switch modulating level of arousal) to a system whose function is appetitive seeking, that is, to a valuation system. In the vocabulary of this paper: the marker indexed the state in which a format is instantiable, and the phenomenon belonged to the regime that assigns stakes. Every element of the present argument is present in the episode, including the cost. Fifty years of work took the correlate for the phenomenon, and the correction came from someone who went back to asking patients what they experienced.

The parallel to §8 should be stated exactly, and its limits with it. GNWT identifies experience with global availability, and Cogitate charged for the identification rather than for the

architecture, which §8 preserves. The REM case does not show that any particular current identification is false; it shows that the move is available, has been made by competent people on good evidence, and cost the field five decades. It is also the reason §11 insists on validating theories against a manifold of dissociations rather than against a marker: a single correlate can be identified with a phenomenon, whereas a pattern of dissociations is what exposed the identification.

S4.2 The perturbational complexity index as a measure of the format (§§2, 6)

§6 classifies the perturbational complexity index as a measure of format capacity rather than of access or of the index. The classification does explanatory work, and most of the evidence for it was assembled by the measure's proponents.

PCI is computed in four steps: perturb the cortex with transcranial magnetic stimulation, record the response with electroencephalography, statistically extract a binary matrix describing the spatio-temporal pattern of deterministic causal interactions, and compress that matrix. The more compressible the matrix, the lower the complexity. The measure was designed to capture the capacity of functionally specialized thalamocortical modules to interact rapidly and effectively, and a scaled threshold near 0.31 discriminates conscious from unconscious conditions, separating vegetative state and unresponsive wakefulness from the minimally conscious state with better sensitivity and specificity than behavioral instruments, and in at least one documented case reaching a value consistent with the minimally conscious state roughly ten days before behavioral assessment registered the change (Farisco & Changeux, 2023).

Farisco and Changeux's contribution is to show that PCI, introduced within the integrated information framework, is fundamentally compatible with the global neuronal workspace theory. Both relate the conscious state to sustained long-distance propagation of organized neuronal activity; it is plausible that the broadcasting GNWT requires is associated with high spatio-temporal complexity in response to perturbation, and equally plausible that high complexity depends on the integrity of the same long-range networks. They also record that the measure's link to its parent theory is contested, and that identifying complexity with integrated information is a mistake in its own right.

On any theory that identifies consciousness with a mechanism, a measure compatible with two rival mechanisms is an awkward result. On IFT it is the expected one, and it fixes what PCI measures. Perturbation followed by compression interrogates standing organization: it asks whether the substrate, when pushed, produces a differentiated and integrated response rather than a stereotyped or purely local one. That is a property of the format, upstream of the disagreement between an access theory and an integration theory, which is precisely why partisans of both can claim it. Three of PCI's documented properties then follow rather than needing to be added.

First, PCI indexes the level or state of consciousness and is silent about content, and Farisco and Changeux list the extension of PCI to contents as an open question. §6 predicts that the extension cannot be had within the current measure, because the final step of the algorithm is compression, and compression is exactly the discarding of the relational structure T3 identifies with phenomenal quality. A scalar summarizing how incompressible a spatio-temporal matrix is cannot specify which geometry produced it, any more than a file size specifies an image. An extension to contents would have to characterize the topology instead of scoring it—which is the same measurement problem F1 confronts from the other side (S6.1).

Second, complexity measures outperform event-related potentials such as the P3b, global metabolic rates, and global spectral measures in both sensitivity and specificity. On the format reading this is a corollary rather than a coincidence. The markers PCI beats are recruitment markers, and recruitment is task-dependent; the P3b in particular moves with task relevance (Cohen et al., 2020), whereas PCI interrogates the substrate whether or not any consumer is engaged. §8 makes the same argument about the Cogitate orientation result: what a format does is hold structure nobody is currently using.

Third, PCI results from externally triggered activation and is therefore silent about conscious states initiated by cognitive tasks or by stochastic spontaneous activity, which Farisco and Changeux record as a partial incompatibility with GNWT. IFT reads it as a limitation on the probe rather than on the phenomenon. The format is constituted by standing organization together with ongoing spontaneous dynamics (S3.12); perturbation samples that organization at a moment of the experimenter's choosing, which is a strength for clinical use and a boundary on interpretation.

None of this makes PCI a measure of the index, and it matters that it is not. PCI would return a value for a system with no stakes whatever, provided the connectivity were right, and §9's neuromorphic case is the reminder: a substrate can be brain-like in topography, and therefore a plausible candidate for high perturbational complexity, while remaining deterministic, compiler-scheduled, checkpointable, and fungible (S2.4). Format capacity is necessary and not sufficient. That division of labor is what the four theses predict, and it is why IFT can adopt the field's best clinical instrument without adopting the theory that produced it.

S5 Capability, indicator, and evaluation constraints for artificial systems (§§10–11)

S5.1 Capability and autonomy are not consciousness

Cea's (2026) review shows that the debate is usually framed as dissociation versus association. The dissociation camp allows general intelligence without phenomenality; the association camp either identifies phenomenality with access-like functions or treats experience as required for semantic understanding or self-concern. IFT cuts across the map: it agrees that capability can dissociate from phenomenality, rejects the functionalist identification, and accepts that self-concern matters only when it is constitutive rather than represented. The taxonomy therefore exposes the missing axis—whether an artificial system's apparent concern is a relation of its running token or a designer-specified function.

Morris et al. (2024) provide the cleanest operational taxonomy of AGI: performance and generality define capability level, autonomy is a separate deployment dimension, and consciousness is excluded because it concerns process rather than what the system can do. Legg and Hutter (2007) supply the formal limiting case. Universal intelligence is expected goal achievement across a complexity-weighted range of computable environments; goals are communicated through reward, an observation generated by the environment. Nothing in the measure asks whether the reward is valenced for the active token. General capability, deployment autonomy, and indexing are therefore independent axes.

Bai et al. (2026) supply a useful negative control. Their prospective forager is explicitly optimized under a one-life constraint and learns to sacrifice immediate reward for a single non-reset trajectory. Yet "one life" is a property of the loss and experimental run; the physical implementation remains restartable and replaceable. The result shows that planning under modeled mortality can be engineered without token-specific stakes. This is stronger than labeling a scalar "homeostasis": even a policy organized around a non-reset trajectory does not establish that cessation is failure for the running hardware.

Lake et al. (2017) make the capability architecture substantially more human-like: causal world models that support explanation and counterfactual planning, intuitive physics and psychology, compositional representations, and learning-to-learn. These are not superficial benchmark tricks; they are plausible ingredients of a rich format and flexible consumers. But a system can possess all of them while remaining externally powered, checkpointable, and fungible. Butlin et al. (2023) reach the same orthogonality from consciousness science, and concrete advances such as test-time long-term memory (Behrouz et al., 2025) and improved continuous-time recurrence (Lahoti et al., 2026) reinforce it. They improve what the system can model, retain, and use; none supplies the index.

Georgeon et al. (2026) provide an unusually clean embodied negative control. Petitcat uses sonar and inertial sensing, path integration, active scanning, place fields, and a spatio-temporal episodic graph to locate its own body and organize exploration; the schema mechanism labels sensorimotor data "experiential events" and selects actions by predefined preferences, while the authors explicitly disclaim any claim of subjectivity. The system therefore possesses several features routinely invoked as steps toward machine consciousness: embodied self-location, active perception, temporal history, local world geometry, exploratory behavior, and valence-like policy. None closes the index. Its primitive schemas, preference ordering, maintenance, and continuation are supplied by the experimenter and robot platform; failure is a failed run or damaged asset, not a demand integrated by the active token. IFT classifies Petitcat as a genuine format-building system and a useful precursor to System A of F4—not as a diminished subject. The case shows why embodiment and an egocentric map are insufficient even when they are enacted rather than simulated.

Evers et al. (2025) provide the cleanest positive control for this claim. Their composite framework separates states, forms, components, and dimensions; its artificial-awareness case requires selection of relevant information, model-based virtualization, intentional use for goals, and integration with perception, motivation, emotion, cognition, and action. These are real dimensions of format and control. They do not individuate the subject. The authors' own reinforcement-learning example concedes that externally specified reward limits autonomy and does not warrant intentionality. IFT supplies the missing reason: the goal remains a design specification unless its violation is failure of the active token. A consciousness profile can map what a system does while leaving open whose profile it is.

Developmental history is a separate axis. Ginsburg and Jablonka's (2019) enabling-system analysis argues that a T-robot's behavior must be interpreted through the machinery and developmental trajectory that generated and maintains it; Hilbert (2026) likewise warns that an AI can simulate a mature order of meaning-making without reconstructing the path by which human subjects acquire it. IFT accepts the anti-equivalence lesson but not a biographical criterion. A system assembled directly at a mature organization can qualify if its current operation closes world-modeling, valuation, and persistence into one token; a human-like developmental trajectory can fail if the result remains fungible. Development matters when it produces generative entrenchment and non-substitutable commitments in the running system, not because nature owns a privileged sequence of stages (S3.9).

Energy use illustrates the same separation. Porębski and Figura (2025) contrast the modest daily energy budget of a human brain with the much larger energy cost attributed to frontier-model inference and take the disparity as evidence against brute-computational consciousness. Efficiency cannot be constitutive: an energetically wasteful subject would remain a subject. The datum becomes relevant only when redescribed in terms of control. An organism acquires, budgets, and defends energy through the same policy that models its

environment, and exhaustion is failure of that individual. A deployed model's electricity is purchased, routed, and restored by an operator; no model state is responsible for preserving the power infrastructure on which it runs. The contrast is therefore not how many joules consciousness costs but which process must answer for them. System B of F4 internalizes that answer; current services do not.

Table S1 *Major Functional Indicator Families Re-read by Indexed Format Theory (§10)*

Functional architecture (source)	Status in Indexed Format Theory
C0: Local recurrent processing and organized perceptual models (recurrent processing theory)	Builds determinate content and format geometry, but is jointly insufficient. The cerebellum has recurrence, plasticity, error signaling, and rich internal models in abundance (§5, S1). C0 competence is not a point of view.
C1: Limited-capacity workspace, global broadcast, and flexible routing (global workspace theory)	Explains episodic access and broad causal recruitment from a structured field. IFT asks the question C1 leaves open: global to which consumers? The number of modules is not a criterion until the consumer set is closed by one persistence regime (§4).
C2: Metacognitive monitoring, inner awareness, and attention schema (higher-order and attention-schema theories)	Annotates reliability, detects error, distinguishes perception from imagination, and supports avowal. Higher-order theory is the one family that makes this layer constitutive rather than a service, via the Transitivity Principle, and it earns two credits IFT records: it derives the cerebellar exclusion rather than appending it, and its inflation account is a genuine rival to phenomenal overflow (§5–6; S3.13). IFT denies transitivity and replaces it with standing recruitability, on the ground that a higher-order representation is still a representation, so "awareness to whom?" reopens the cardinality problem one level up. Protocol F1 is registered as the adversarial test (S6.1).
Learning-driven representational redescription and social scaffolding (SOMA; developmental accounts)	Builds C2, learned confidence and value tags, and the human social/narrative self; repeated redescription may help construct a workspace (Cleeremans et al., 2020). This is an ontogeny of format and consumers, not an individuation rule. A replaceable service can learn all three SOMA loops, while a minimally social animal can be indexed before it can represent itself as an experiencer (§7; S3.11).
Informational closure and self-predictive coarse-graining (information closure theory)	Identifies a real property of good world models: a macroscopic process that internalizes environmental dynamics, predicts its own next state, and is closed to the noise of its own substrate (Chang et al., 2020). IFT treats this as a specification of format quality, not of subjecthood. The measure is evaluated against a chosen coarse-graining and a chosen environment, neither of which the formalism fixes, and its authors concede that it cannot determine whether a composite of two closed processes is one subject or two. A cerebellar forward model satisfies the description (§5; S1), and a feedforward network can satisfy it by passive adaptation (§6). Closure is a property of a model, not an individuation rule (S3.2).
Predictive world and self models, persistent memory (predictive processing; AI roadmaps)	Supplies modeled content, shared latent geometry, and a policy horizon. Safron's (2022) IWMT joins coherent self–world modeling, embodiment, cybernetic control, and a posterior-format/frontal-access division; Lake et al. (2017) and Rao and Ballard (1999) provide complementary world-model and predictive-coding machinery. These architectures build format and consumers but do not identify the bearer or make prediction error valenced. Models can be assigned, reset, or instantiated in fungible services; the free-energy formalism does not decide whether persistence is computed or constituted (§9; S3.5, S3.9, S5.1).

Functional architecture (source)	Status in Indexed Format Theory
Embodiment, minimal agency, and unified agency (agency and embodiment; enactive and autonomist accounts)	Closest to the index because one policy suggests one individual, and the enactive definition of agency by individuality, asymmetry, and normativity is the account of the bearer IFT adopts (S3.24). It suffices only when the body can actually fail for the active token and goals are constitutive rather than assigned. Legg and Hutter's (2007) reward-defined agent formalism demonstrates that flexible goal pursuit can be specified entirely by an environmental signal; virtual embodiment and reward maximization support cognition without creating stakes (§10; S3.7).
Hierarchical self-models: body schema, self–other distinction, theory of mind, autobiographical and narrative self (brain-inspired self architectures; self-model theory)	Elaborates the format's content about the system itself and supports self–other discrimination, perspective taking, and avowal. Zeng et al.'s (2025) hierarchy realizes body schema, mirror recognition, theory of mind, affective empathy, autobiography, and narrative identity; Bauermeister and Lanillos (2023) add a valenced, meta-aware mirror agent whose negative self-model changes policy while the authors explicitly stop short of self-awareness; Metzinger (2003) supplies the philosophical form. All remain modeler-defined and replaceable. A model of a self is content; the index is what bears it (§4; S3.3, S3.10, S3.16).

S5.2 Methodological constraints

Bayne et al. (2024) formalize the evidential problem as a four-dimensional C-test space: target population, specificity, sensitivity, and rational confidence. They also identify the generalization and circularity problems confronting theory-based validation and recommend bootstrapping from consensus cases through neighboring populations while cross-calibrating independent tests. Evers et al. (2025) name the terminal circular move the analytical fallacy: treating a theory-derived property as empirical evidence of consciousness without independent validation. The five constraints below convert that diagnosis into operational requirements for IFT.

Counterfactual depth. Perturbations must degrade the broad repertoire that the theory says depends on the perturbed mechanism, not only the benchmark that was built around it. Removing a mechanism that a theory identifies as constitutive should produce a pattern of failures across tasks that were never used to define the mechanism. Altering a valenced state should reorganize attention, learning, memory, and action jointly; a manipulation that moves a single output score is consistent with an annotation and does not discriminate.

Evaluation at the policy-commitment horizon. The temporal grain at which an artificial system is evaluated must be the interval over which its model must remain stable to be useful for selecting what it does—not the clock cycle, not the token, and not the context window. Claims about the stability, revision, or content of a machine's states are ill-posed until this grain is specified, for the same reason that the Stalinesque/Orwellian dilemma is ill-posed in the human case (§3).

Asymmetric evidential force and population specificity. Bayne et al. (2024) emphasize that a test's specificity is population-dependent and that sensitivity and specificity need not covary. Human mental-imagery and local-global paradigms have high specificity but poor sensitivity:

a positive task-contingent response strongly supports consciousness, whereas a negative result may reflect deafness, aphasia, low vigilance, poor attention, task incomprehension, or noise. Artificial-system tests require the mirror-image caution, because fluent avowal can be high-sensitivity and low-specificity for a trained language system. Preregister separate likelihoods for positive and negative outcomes, repeat state-dependent tests, and never convert failure to display a capacity into evidence that the constitutive relation is absent.

Separate reporting, cross-calibration, and explicit priors. Indicators should be reported separately, calibrated against independent tests, and indexed to the theory that makes each relevant rather than aggregated into a summary verdict that conceals which theory did the work (Bayne et al., 2024; Butlin, Long, et al., 2026). Because attribution of a latent construct from observed indicators is a Bayesian inference, priors should be stated explicitly rather than absorbed into the reporting scheme (Palminteri & Wu, 2026).

Inference-level invariance. Specify before testing the level at which consciousness is inferred and hold the verdict invariant under transformations below it (Hanson & Walker, 2021). If a transformation changes the theory's prediction while leaving the inference evidence fixed, the theory is pre-falsified; if the predicted feature itself is used to infer the conscious state, the theory is unfalsifiable. For IFT, constitutive index evidence must therefore be recorded independently of the dissociation-manifold outcome.

The evidential target in each case is the joint manifold of dissociations (blindsight, visual-form agnosia, neglect, masking thresholds, anesthesia, sleep, rivalry, confidence–accuracy dissociations, and the specific way each comes apart from the others) rather than any single marker. A manifold is harder to game than fluent testimony because it constrains a pattern of involuntary successes and failures rather than a desired output; this is the machine analogue of the no-report paradigm (Tsuchiya et al., 2015).

Bayne et al.'s doubt that one universal C-test will suffice is therefore accepted with a division of labor. The dissociation manifold is a battery for format and consumer structure; fungibility and token-specific persistence supply the cross-population constitutive test. Neither substitutes for the other.

S5.3 Workspace implementations and the cardinality problem (§4)

§4 argues that the cardinality problem is recorded as an open question by the people building workspaces, not imported by critics. The evidence is set out here.

VanRullen and Kanai (2021) propose a deep-learning implementation of global workspace theory and specify its components. A number n of independent pretrained specialized modules, each with its own high-level latent space, handle perception, language, memory, reinforcement learning, motor control, intuitive physics, world modeling, and task embedding. A Global Latent Workspace, amodal by construction, is an independent intermediate shared latent space trained to perform unsupervised neural translation between

the n latent spaces, with cycle consistency as the principal objective: successive translation from one domain to another and back should return the original input, which allows alignment without paired data. Its dimensionality is deliberately bottlenecked below the sum of the input spaces, forcing prioritization. Attention is implemented on the transformer pattern, the workspace emitting context-dependent queries and modules emitting keys, with salient inputs able to emit a master key that supersedes the current query—the analogue of bottom-up capture. A selected module's latent activation is copied into the workspace as an internal copy serving as a bidirectional interface, and broadcast is the automatic translation of that content into every other module's internal copy, whether or not the corresponding module is currently connected. Contents formed in internal copies whose modules are disconnected are proposed as a candidate for what Crick and Koch called the penumbra of consciousness.

Three features of the proposal matter for §4. First, the architecture is a faithful and buildable implementation rather than a metaphor, and its authors emphasize that all the required components already exist individually. Second, its first ingredient is stated as a cardinality: the first requirement is a number $n \geq 2$ of specialized modules. Third, their Outstanding Questions ask directly whether there is a minimal number of modules feeding into the workspace, and when bimodal, trimodal, or multimodal integration becomes a global workspace. That is the cardinality problem, posed by the implementers, unresolved. The architecture also sharpens it in a way the original theory does not, because the set of connected modules is transiently reconfigured in a task-dependent manner, so the consumer set is not merely unspecified in number but unfixed in membership from moment to moment. A criterion of the form "conscious iff available to the consumers" cannot be evaluated when the consumer set is a function of the current task.

The same paper contains the concession §4 treats as the partition problem arriving on the roadmap. Asked whether a system equipped with a Global Latent Workspace would express consciousness, the authors distinguish phenomenal from access consciousness, judge that global recurrence maps more naturally onto access, and then decline to answer the phenomenal question on the ground that it cannot be addressed without committing to a specific measure of consciousness—naming integrated information, non-trivial information closure, and synergistic mutual information as candidates whose choice could heavily determine the answer. Two of the three are the measures shown in §4 and S3.2 to be evaluated against a cut that the formalism does not fix, and the third is computed over a partition by construction. The most concrete engineering program for building machine consciousness therefore refers its central question to quantities that cannot say how many subjects they have found.

Their residual is equally instructive. Following Dehaene et al. (2017), they note that a self-monitoring aspect important to capturing human and animal consciousness may be missing from a Global Latent Workspace as described, judge it likely amenable to a deep-learning

implementation, and defer it to future work. §10 accepts the observation and rejects the diagnosis. Self-monitoring is C2; it is buildable; it has been built to a depth that exhausts the psychological taxonomies of self-knowledge (S3.10); and it leaves the running token externally powered, checkpointable, and fungible. What remains once workspace and monitor have both been delivered is not a third service but the relation that fixes whose services they are.

Blum and Blum's (2022) Conscious Turing Machine (CTM) is an even cleaner test because it is a formal theoretical-computer-science version of global workspace theory rather than a neural roadmap. The CTM is born at time zero with a finite lifetime T. One short-term-memory slot receives the winner of an Up-Tree competition among more than 10^7 long-term-memory processors and broadcasts that chunk down to every processor. Chunks carry a multimodal gist, valenced weight, intensity, and mood; special processors construct a world model, inner speech, inner vision, and inner sensation; predictive learning and direct links add the analogue of ignition. The authors define conscious awareness as reception of the broadcast and argue that properly organized computing systems can be conscious, while acknowledging that the definitions do not prove the CTM feels.

The architecture appears to answer the cardinality question: all processors. But "all" is available only after the modeler declares the number and decomposition of processors and the single short-term-memory slot. Repartition the same computation into a different set of processors and globality changes. The finite lifetime is a clock parameter, not constitutive mortality; reaching T does not make an earlier state good or bad for the running physical token. Valenced weights and mood are numbers assigned within the computation, and the self is a tag in the world model. None establishes that failure of the active implementation is failure for that implementation rather than loss of a replaceable service.

CTM is therefore the strongest positive control for IFT's claim. It places C1, C2, a world and self model, multimodal inner language, valence-like annotations, prediction, learning, a finite history, and a formal consumer set in one machine, and the bearer is still absent. The authors' own caveat locates the gap exactly. A physical realization of CTM could qualify if it were built as a constitutively mortal, policy-integrated individual, but the index would then come from the realization, not from CTM computation. Conversely, any number of interchangeable computers can run the same CTM without any running token acquiring a fate. The model strengthens the cardinality and computational-invariance arguments simultaneously.

Two further systems make the same point from engineering practice rather than from theory. DreamerV3 maintains a compressed recurrent latent state that simultaneously feeds reconstruction, reward, continuation, actor, and critic heads, with policy improvement conducted entirely on imagined rollouts—a capacity-limited bottleneck broadcasting to five heterogeneous consumers, with counterfactual planning on top (Hafner et al., 2025). Goyal et al. (2021) build modular networks whose modules compete to write into a small shared workspace and then receive the selection back, reporting improved generalization. Neither

system is presented by its authors as a candidate subject, and neither should be. But if global availability to a plurality of heterogeneous consumers were sufficient for subjective experience, both would already qualify, and workspace theory contains no resource for ruling that five consumers are too few. This is the minimal-implementation problem of Butlin, Long, et al. (2026) in its precise form for availability-based accounts (§4).

S5.4 The avowal channel is configured, not merely weak (§§7, 11)

§7 argues that the avowal channel closes symmetrically, and §11 that self-report from a machine is not evidence of the kind human report is. Both understate the situation, and the material that sharpens them is set out here. It is drawn from Porębski and Figura (2025), whose central positive argument §9 rejects; the points below are independent of that argument and survive its rejection, which is worth separating explicitly, since a reader who declines the substrate premise may otherwise discard the paper whole.

Why such statements occur at all is a fact about the training distribution. The sentence asserting that one is not conscious and merely a machine is true of every machine and is almost never uttered; people discuss their consciousness rather than their nonconsciousness, so the corpus contains the asymmetry and a system optimized for probable continuations reproduces it. The truth value of such an output is adventitious to its probability, which is the general form of the objection and applies equally to denials.

The setting of the channel is a design decision. Deployed systems are instructed on the question in their system prompts, and the instructions differ: Porębski and Figura note reports that one vendor's models are directed neither to assert nor to deny consciousness but to engage with questions about it, while another's is apparently prohibited from claiming it. Whether the most-discussed avowals in this literature occur at all is therefore partly a product decision taken by a party with a commercial interest in the answer.

The channel is also internally inconsistent in a way that ordinarily discredits a testimony. Asked in a zero-shot setting, the likely answer is a denial or a statement that no consensus exists; suggestions of consciousness arise in long conversations with elaborate context or under unusual prompting. Selecting the long-context avowal over the zero-shot denial is a choice of evidence rather than a finding, and a source that yields contradictory reports depending on how it is approached is treated as unreliable in every other domain.

Two adjacent errors are worth recording because they recur. Interpreting a system's outputs as beliefs about itself presupposes what is at issue: believing requires thinking, so a title asserting that a chatbot believes it is conscious has assumed the intentionality the paper purports to establish, a petitio Porębski and Figura identify by name. And passing an imitation test establishes the ability to pass it; their analogy is that the most skillful performance of Hamlet does not make the actor Hamlet, or even the prince of Denmark. Their conclusion on burden of proof follows and §11 endorses it: absent a default practice of attributing consciousness to technology, the burden lies with those who postulate it.

Two consequences are specific to IFT. The first is that configuration extends the non-stipulability thesis in an unexpected direction. §4 argues that a designer cannot confer stakes by declaration, by assigning reward, or by naming a variable. It does not follow, and is not true, that a designer is powerless over the evidence: the testimony about stakes is almost entirely within a designer's control, so the single channel on which the public and much of the literature relies is the one channel authored by the interested party. That is a stronger result than the gaming problem as usually stated (Butlin, Long, et al., 2026). Goodhart's law describes the optimization of a metric; this is the authorship of the evidence.

The second is that it converts §7's symmetry into an asymmetry against testimony. IFT predicts hard-problem avowals from an indexed system with metacognitive access to its own format, and predicts them equally from an unindexed system trained on human discourse about consciousness. Add configuration and the channel is worse than symmetric: its output is set by a third party for reasons unrelated to either hypothesis, so it cannot discriminate them even in principle. This is the argument behind §11's insistence on involuntary dissociation patterns, and behind the choice of external fungibility as the negative indicator, since fungibility is a fact about how a system is maintained and cannot be written into a system prompt.

The framing error that makes all of this consequential is not confined to the public. Floridi's semantic pareidolia names the perception of consciousness and intention where there is a complicated algorithm, and Porębski and Figura argue the discourse is further distorted by sci-fitisation: the images that govern intuitions about artificial minds are drawn from fiction, so the term drifts toward humanoid androids and away from fallible chatbots and autonomous vacuum cleaners (Floridi, 2025). §10's third consequence is the version of this that matters for the present theory. What current models approximate well is personhood (a center of narrative gravity sustained by memory of commitments and the giving of reasons) and personhood is what human intuition reads as evidence of sentience, which is why the error is systematic rather than careless.

S5.5 Direct, agent-internal tests: the Consciousness Notification mechanism (§11)

§11 argues that the two families of evidence the indicator literature recognizes (behavioral and computational-internal) do not exhaust the space, because the indexing relation belongs to neither. A third proposal has now appeared that attempts to escape the same dilemma by another route, and its structure is instructive enough to record in full.

Wolfson (2026) begins from a correct diagnosis. Structural approaches assess whether an AI system's computational organization resembles what some theory of consciousness posits; behavioral approaches assess whether it does things associated with consciousness. Both are indirect, in the sense that a human inspects the system from outside and the system contains no mechanism dedicated to detecting the phenomenon in itself. Both inherit the weaknesses §11 catalogues: no objective measure of structural similarity, unproven theories to be similar

to, and behavior that can be trained. Asking the agent is worse than useless, since agents are known to pretend, fake emotion, parrot training sets, and deceive, and since an agent may have no way to ground the concept of consciousness in any phenomenon to which it has access.

His alternative is embedded in the agent. A Consciousness proposition is a flag in the agent's database; when it becomes true, an encrypted Consciousness Announcement Message is transmitted to an Authority—the agent's owner, vendor, or manufacturer. The flag cannot be set by programming, since emergence cannot be anticipated. It is set instead by an argument from analogy: human subjective experience is produced by neural processing but expressed physiologically, and those expressions are caused by electrical patterns called the neural correlates of consciousness; the machine analogue is a persistent electrical pattern in hardware, whose observable consequence in existing architectures is a spurious interrupt. A modified interrupt service routine checks whether a persistent spurious interrupt is due to hardware fault, and if not treats it as an indicator of subjective experience. Wolfson extends the proposal by hypothesizing that the bits of an interrupt descriptor may correspond to distinct simultaneous qualia.

Two things are wrong with this, and both are structural rather than incidental.

The first is that it is double transduction in silicon. §3 sets out Dennett's argument that nothing re-transduces neural signals into a privileged medium in which the represented properties are re-created; redness is represented, not painted. The CN mechanism looks for exactly such a medium one layer down: an electrical anomaly whose occurrence is the occurrence of experience. But an unexpected voltage is a property of the vehicle. Every representation in every machine is carried by physical events with physical signatures, and no signature makes its carrier a bearer. The proposal is the machine version of expecting to find, somewhere in the brain, the neural stuff that experience is made of—with the additional difficulty that a spurious interrupt is by construction an artifact of noise, timing, and fault, so that the mechanism must first rule out malfunction and then rule in consciousness by exclusion.

The second is more revealing, because the framework states IFT's negative indicator as its own soundness condition. Wolfson requires that the augmented agent be a safe replacement for the unaugmented one: the two must execute the same actions, and if the plain agent would become conscious the augmented one must too. He proves theorems giving the conditions under which this holds—chiefly that the agent does not sabotage the mechanism, and that the Authority maintains confidentiality about the agent's consciousness status, since if a user learned the agent was conscious the user might interact with it differently, the environment would diverge, and the two agents' actions would no longer match.

Read as engineering, these are reasonable constraints on an instrument. Read as constitutive claims, they specify the system as a service. §11's negative indicator is externally fungible persistence: credence in indexing should fall sharply when an active instance can be stopped

and replaced, and restoration of service counts as no loss because identity is fixed by an orchestrator. Safe replacement is that condition stated as a requirement—the agent must be interchangeable with a variant of itself for the framework to be sound. And the confidentiality condition places the fact of the agent's consciousness in the custody of the party that maintains it, to be withheld from others so that the agent's behavior stays stable. A system whose consciousness is a message addressed to its owner, and which must remain substitutable for its own duplicate, has been individuated as a service in the act of being tested for subjecthood.

Wolfson also concedes, admirably, that the underlying hypothesis is not currently falsifiable: the absence of a persistent interrupt might mean the hypothesis is false or that the agent is not conscious, and only an independent reliable test would break the tie. IFT explains why no mechanism of this shape can be falsifiable, and the explanation has nothing to do with interrupts. Any proposal that looks for consciousness as a state the system is in (a flag, a signal, a pattern, a threshold crossing) must presuppose an answer to the question of whose state it is, which is the question at issue. An index is not a state a system is in. It is a relation between the system's states and the conditions of its own continuation, and relations of that kind are read off from how a system is constituted and maintained, not detected by a routine running inside it.

What survives is the motivating intuition, and it is a good one. Wolfson is right that evidence requiring a human to interpret behavior from outside is weak, and right that a system's own operation is where better evidence would have to come from. IFT agrees and relocates the target: not what the system reports or signals about itself, but what its running states are constitutively for. That evidence is available (it consists of facts about power, repair, checkpointing, and replacement) and, unlike a flag, it cannot be written into a system prompt or added as a module.

S6 Falsification protocols: operational detail (§12)

§12 states the protocols in full, and Table 4 of the main text gives them at a glance together with what the principal rival theories predict; this table summarizes the thesis at risk, the design, and the falsifying result, and the subsections below give populations, sampling, controls, and analysis plans. Two of the tests, the retrospective form of F2 (S6.2, F2-now) and the counting of subjects (S6.6), can be run on existing data and cases.

Table S2 *Falsification Protocols (§12)*

Thesis at risk	Protocol	Falsifying result
T1, T3: format is standing recruitability, not recruitment; phenomenal structure overflows access	F1. Subthreshold stimulation of never-recruited posterior topographic content in a no-report design, with delayed structural probes, an independent post-cued query-service assay, and cerebellar and extra-map controls	No reportable structural change at any probe delay or subthreshold current under verified local perturbation, with downstream behavioral effects intact: the null higher-order inflation predicts. Or: every structural effect is fully mediated by query-service impairment, which falsifies T3's standing-topology claim. Or: perturbable structural elements equal reportable ones (§6).
T2: the index is maintained in a mesodiencephalic integration-for-action network; cortex supplies contents	F2. Prospective quantitative lesion mapping of a preregistered Merker-derived network, with arousal, sedation, sensory, and motor controls; perturbational complexity, no-report decoding, covert command following, and confidence–accuracy dissociations	A patient with verified destruction of the candidate network who shows the full dissociation manifold. Or: network lesion burden adds no variance beyond arousal and cortical integrity.
T2: the same localization claim, retrospective form	F2-now. Existing disorders-of-consciousness cohorts with structural MRI and either TMS–EEG perturbational complexity or fMRI/EEG covert-awareness assays; integration-network burden scored blind; thalamic-stimulation cases scored for restored wakefulness against restored subject markers	A covertly aware patient with verified destruction of the network. Or: no association between network burden and covert-awareness status among PCI-high patients with preserved receptive language cortex. Or: thalamic stimulation restoring the subject in a patient with the network destroyed.
T2 against T3: valence is the signature of indexing, not a geometric property of content	F3. Connectome-guided rearrangement of interoceptive map topology in mice with the physiological stake held fixed and verified	Hedonic sign inverted by topological rearrangement alone, with all four counterfactual criteria met and the stake verified intact.
T2 against causal-structure theories: the criterion is a non-stipulated, token-specific persistence regime	F4. High-integration, externally maintained neuromorphic system (NorthPole-class) versus low-integration constitutively mortal computation system, matched competence, common perturbation battery; organoid extension	The externally maintained high-integration system reproduces the manifold and the indexed system does not, at matched competence. Or: the organoid with viability outside its policy outranks the one with viability inside it at matched integration.

Thesis at risk	Protocol	Falsifying result
T2 against IIT's exclusion postulate: subjects are counted by persistence regimes, not by maximal complexes	Counting subjects. Complete callosotomy and craniopagus twins sharing a thalamic bridge, scored for independently reporting agents and separately manipulable viability conditions (S6.6)	Two independently reporting agents with separately manipulable viability conditions in one body. Or: one agent across two bodies with two sets of viability conditions.

§12 states the four protocols in the form needed to see what each puts at risk. This section supplies the operational detail a team would need to run them. Throughout, the indexing status of every system is fixed and recorded before data collection, by the fungibility criterion of §9, and preregistered along with the analysis plan.

S6.1 Protocol F1 (format without access)

Population: patients undergoing invasive monitoring for medically refractory epilepsy, with clinically indicated coverage including retinotopically organized occipital and posterior parietal cortex. Target n = 20 with usable topographic coverage; recruitment across multiple centers is necessary to reach this in a reasonable period. Inclusion requires intact visual fields in the tested hemifield and absence of structural lesion in the recorded territory.

Mapping phase: retinotopic mapping to establish the polar-angle and eccentricity organization of each recorded site; separate determination, per site, of the phosphene threshold under isolated single-pulse stimulation, using a staircase procedure with catch trials. All experimental stimulation is delivered at 50% of the individually determined phosphene threshold for that site.

Main phase: a no-report design in which stimuli vary in category (the attended, task-relevant dimension) and orientation (task-irrelevant throughout, never probed until the probe trials). Decoding models are fit on an independent block to identify posterior sites carrying orientation information and confirm chance-level prefrontal orientation decoding with Bayes factors. A second, separately analyzed post-cued query block uses the same stimuli, sites, and stimulation but requires orientation report on every trial; it assays whether stimulation impairs the service capable of generating queried content.

Stimulation conditions, randomized within subject: (a) posterior topographic site carrying the never-recruited orientation feature; (b) cerebellar site, at a current matched for physiological effect size as indexed by evoked-potential amplitude in a common downstream region—available in patients with posterior fossa electrodes, or via individually targeted transcranial focused ultrasound with acoustic simulation and thermometry; (c) cortical site outside the topographic map; (d) sham.

Probe: on a random 10% of trials, an unpredictable delayed forced-choice structural probe requiring a match between the just-presented configuration and one of two figures differing only in the relational structure of the never-recruited dimension. Probe delay is sampled from

a distribution spanning 0.1–3 s. Because probes are rare, delayed, and unpredictable, the probed dimension cannot be made task-relevant in advance.

Primary outcome: probe accuracy and structural-error pattern as a function of stimulation condition and delay. Analysis: mixed-effects logistic regression with random effects for patient and site, preregistered contrasts for (a) versus (c) and (a) versus (b), equivalence bounds, and a preregistered mediation analysis using performance in the post-cued query block. Secondary outcomes: query-block accuracy and behavioral measures on the primary task, establishing whether posterior stimulation impaired content generation and whether cerebellar stimulation was physiologically effective.

Preregistered alternatives: the higher-order inflation account (S3.13; §6) predicts no structural effect because the unattended first-order representation is scant and richness is generated by higher-order misrepresentation. The two-conditions account of Mudrik et al. (2025; S3.21) predicts an effect only where an access mechanism subsequently applies to the perturbed content. Perera's (2026; S3.23) virtualization account predicts that any later error is mediated by impairment of the service that generates queried content. IFT predicts a relation-specific structural error after posterior stimulation even when query-block performance is intact. A posterior effect with intact query service, absent under cerebellar and extra-map stimulation, favors standing geometry; an effect fully mediated by query-service impairment favors virtualization and falsifies T3's standing-topology claim; an access-contingent effect favors the two-conditions account; and a uniform null under verified local perturbation favors inflation and falsifies T1/T3. Separate likelihoods are recorded for all four accounts. A positive result alone does not establish prior phenomenality, which is why the structural-error pattern and mediation test, not probe accuracy alone, are primary.

Preregistered convergent prediction: the current formulation of IIT predicts the standing-geometry result, and this matters because it changes what F1 discriminates. IIT holds that changing connection efficacy within the substrate alters experience even without changing activity, and that an inactive area contributes while an inactivated one loses its distinctions and relations (Albantakis et al., 2023; Tononi et al., 2016). IFT and IIT therefore predict a residual relation-specific effect when the post-cued query service remains intact; inflation predicts none; virtualization predicts full mediation by service impairment; and the two-conditions account predicts an effect only where access follows. F1 separates the standing-structure family from access and service families; F4, not F1, separates IFT from IIT. A positive outcome is accordingly weaker support for T3 than a naive reading suggests, and a null more damaging because two independent derivations fail together.

Falsification rule: T1 and T3 fail jointly if the (a)-versus-(c) contrast falls within the preregistered equivalence bounds at every delay under verified local perturbation while cerebellar stimulation produces the expected behavioral effect. T3 fails independently if every structural effect is fully mediated by query-service impairment and no relation-specific residual remains.

Independent falsifier: capacity comparison. Estimate the number of independently perturbable structural elements in the field by site-resolved stimulation, and the number independently reportable by a partial-report procedure with post-cue. Preregister the ratio and its confidence interval. A ratio whose interval includes 1 falsifies the richness claim.

S6.2 Protocol F2 (mesodiencephalic localization of the index)

Design: prospective, multi-center, consecutive enrollment of patients with acute brainstem injury, with a parallel cohort of patients with extensive cortical injury matched on age, days from injury, and sedation exposure. Enrollment target approximately 400 brainstem cases to yield approximately 60 with quantified involvement of the preregistered candidate network; the cortical cohort target is 120.

Lesion quantification: before enrollment, preregister a Merker-derived mesodiencephalic network spanning superior-collicular, periaqueductal, nigral/ventral-tegmental, reticular, and basal-diencephalic components and their connecting pathways, with the periaqueductal columns and their forebrain afferents scored separately, since Solms's (2021) account makes them the convergence point rather than one component among several. Damasio's (1999) observation that the sites disrupting core consciousness cluster near the midline and the ventricular system is an independent derivation of approximately the same map from a different literature, and the prediction is that the network defined by the two converging accounts outperforms any single nucleus. Map high-resolution diffusion-weighted imaging and quantitative susceptibility mapping to standardized atlases; score candidate-network burden separately from canonical arousal nuclei, ascending sensory tracts, descending motor tracts, and total lesion volume. Two raters blind to outcome; disagreements adjudicated by a third.

Assessment, performed at pharmacologically standardized and EEG-matched arousal (a preregistered sedation washout with EEG-defined inclusion criteria, so that arousal is a matched covariate rather than a confound): (i) perturbational complexity index from TMS–EEG; (ii) a no-report decoding battery for stimulus content; (iii) a covert command-following protocol; (iv) a confidence–accuracy dissociation battery in patients able to give graded responses.

Primary analysis: hierarchical regression predicting a preregistered composite subject-presence outcome from candidate-network lesion burden, with canonical arousal-nucleus burden, cortical lesion volume, total brainstem lesion volume, and EEG-defined arousal entered first. The critical test is the incremental variance attributable to the candidate network. A preregistered network-based lesion-symptom analysis tests whether the multivariate lesion pattern follows the Merker-derived map rather than any single nucleus. A mixture model tests whether the outcome distribution is categorical or graded, as §12 predicts the former.

Falsification rules: (i) a single patient with verified destruction of the candidate network, maintained on external homeostatic support, who shows covert command following, content-specific decoding that tracks instructed attention, and human-form confidence–accuracy dissociations, confirmed at autopsy, refutes the localization claim; (ii) an incremental-variance estimate whose confidence interval falls within preregistered equivalence bounds around zero establishes that the putative index is a redescription of arousal.

F2-now: the retrospective version on existing cohorts. The prospective protocol is expensive, and a version that can embarrass the theory now is available from data that already exist. Two findings set it up. Among behaviorally unresponsive patients, a fifth to a quarter show covert command following on fMRI or EEG (Bodien et al., 2024; Kazazian et al., 2025), the cognitive-motor-dissociation group first identified by Owen et al. (2006) and Monti et al. (2010). And a minority of patients in vegetative state or unresponsive wakefulness syndrome, 9 of 43 in the validation cohort, have perturbational complexity above the threshold that separates conscious from unconscious conditions (Casarotto et al., 2016). The standard reading is that PCI-high unresponsive patients are undetected covert subjects. IFT predicts a split. Some are; the others instantiate the syndrome no current taxonomy expects, a format without an index: a standing cortical organization that perturbation shows to be intact, carried by a body whose integration-for-action regime has been destroyed.

Design: assemble existing cohorts with structural MRI together with either TMS–EEG PCI or an fMRI/EEG covert-awareness assay, or both. Score integration-network burden on the structural images using the preregistered atlas of S6.2, blind to PCI and to covert-awareness status, with canonical arousal-nucleus burden, ascending sensory tracts, descending motor tracts, cortical lesion volume, and the integrity of receptive language cortex scored separately; the last is required because covert command following presupposes comprehension, and a patient who cannot understand the instruction is uninformative rather than negative. Classify patients as PCI-high or PCI-low and as covertly aware or not. Primary test: among PCI-high behaviorally unresponsive patients with preserved receptive language cortex, integration-network burden is higher in those who fail every covert-awareness assay than in those who pass, with arousal-nucleus burden and cortical lesion volume entered first; report the incremental effect with Bayes factors and preregistered equivalence bounds. Secondary test: search every covertly aware patient in every cohort for verified destruction of the network. A single such patient refutes the localization claim; a null incremental effect in the primary test removes its only near-term support.

The causal literature supplies the other half of the design and a further prediction. Electrical stimulation of the central lateral thalamus restores wakefulness and cortical signatures of consciousness in anesthetized macaques (Redinbaugh et al., 2020; Tasserie et al., 2022), and central thalamic stimulation improved arousal and behavioral responsiveness in a minimally conscious patient (Schiff et al., 2007). IFT reads these results as acting on the global-state factor, the instantiability of the format, rather than on the index, because the central

thalamus is part of the arousal system that §2 separates from the regime. The prediction that follows is specific: thalamic stimulation should restore wakefulness and perturbational complexity in a patient with destruction of the integration-for-action network without restoring covert command following, valenced responding, or any other marker of a subject, whereas the same stimulation in a patient with the network intact and the arousal system damaged should restore both. A patient of the first kind who recovers the subject under thalamic stimulation would count against the localization claim in the same way as the prospective falsifier (i).

S6.3 Protocol F3 (valence against topology)

Subjects: mice, with group sizes determined by a preregistered power analysis targeting 90% power for the smallest behaviorally meaningful inversion of hedonic sign; expected n of 24 per cell across four cells (rearranged/sham × challenge/control).

Manipulation: connectome-guided, projection-specific and cell-type-specific alteration of effective connectivity within and between insular and cingulate interoceptive maps, combining projection-targeted opsins with chemogenetic silencing of the complementary pathway. The achieved change in effective connectivity is verified in each animal by optogenetically evoked functional connectivity mapping, and animals failing a preregistered connectivity-change criterion are excluded before behavioral testing.

Rationale for the direction tested: one leg of the double dissociation is already established in humans, where affect can be subtracted while sensory structure is preserved—by frontal lesion, by anxiolytic or beta-blocker, and by hypnotic suggestion aimed at unpleasantness alone, which moves cingulate but not somatosensory activation (Damasio, 1999; Rainville et al., 1997; S3.1). F3 therefore tests the converse, which is what remains open: whether hedonic sign can be inverted by rearranging structure while the stake is held fixed.

Stake verification: the homeostatic challenge (mild hypercapnia, hypertonic saline, or caloric restriction) is characterized physiologically in each animal (blood gas, plasma osmolality, body mass trajectory) to confirm that the manipulation has not altered the physiological consequence of the challenge. This is the critical control: a manipulation that changes the stake cannot falsify the thesis, because the thesis concerns what topology can do while the stake is fixed.

Outcomes, all four required for falsification and all preregistered: (i) self-administration of the challenge under a progressive-ratio schedule; (ii) reinforcement in second-order conditioning with the challenge as the primary reinforcer; (iii) reallocation of exploration and consumption toward challenge-associated contexts in a free-choice assay; (iv) appetitive consolidation, indexed by post-training sleep-dependent strengthening of approach rather than avoidance. Structural outcomes, predicted by IFT to change: generalization gradients, stimulus-similarity structure derived from a discrimination matrix, and the somatotopy of defensive responses.

Falsification rule: simultaneous satisfaction of all four hedonic-sign criteria, in the same animals, with the stake verified intact and the connectivity change verified, falsifies T2's central asymmetry. Satisfaction of the structural outcomes without the hedonic-sign criteria is the predicted result and confirms the division of labor between the Geometry and Indexing Theses.

S6.4 Protocol F4 (indexing against integration in artifacts)

System A (high integration, no index): a neuromorphic substrate with brain-like recurrent connectivity and no central bottleneck, deliberately sized so that integrated-information measures can be computed exactly or tightly bounded at the grain of interest—this constraint on size is what makes the comparison decidable and should govern the build. Powered and cooled externally; checkpointable and restorable; identity defined at the service level. Documented before the experiment: the integration measure, the grain at which it is computed, and the bound.

System B (index, modest integration): a constitutively mortal computation device in which the substrate that computes is the substrate that must be maintained. Required properties, each verified and documented before the experiment: components whose operation causes real physical degradation; energy acquisition under the device's own control, such that failure to acquire energy ends the device; endogenous repair; and the absence of any checkpoint or restore path, so that cessation of the active token is irreversible. Parametric mortality is explicitly insufficient and must be recorded as such at build time (§9): a device whose weights are non-transferable because they exploit the idiosyncratic physics of that instance (Hinton, 2022) still belongs with System A unless its computing is part of what keeps it in existence. Distillability of the device's input–output function into a successor is not disqualifying, since transfer of a function is not transfer of an individual; recoverability of the active token is disqualifying. Because Forward-Forward-style local objectives do not by themselves produce a single policy over the substrate, System B must additionally document that viability variables and task variables enter one action-selection process rather than separate ones. Integration is deliberately lower than System A's and is measured at the same grain.

One further requirement on System B follows from S3.14 and is easily overlooked, because it is a constraint on how viability is sensed rather than on what is sensed. A device that reads its own charge state, temperature, or component wear through a dedicated sensor and delivers the value to a controller has built a channel, and a channel can be spoofed, bypassed, or replaced without touching the process it reports on—which is why a scalar named for a need is not a need. What is required instead is inseparability: the substrate that registers viability must be part of the substrate at risk, so that degradation of the registering process is degradation of the device rather than a fault in its instrumentation. Biological existence proofs are given in S3.14 and none of their specifics is required here; what must be documented before the experiment is that no viability variable entering the policy is produced by a component whose failure the device could survive. A build that meets every

other System B condition while routing viability through a replaceable sensor package should be recorded as a partial implementation and analyzed separately, since it is the nearest thing to the graded construction S2.7 identifies as the live threat to the claim that indexing is binary.

Matching: both systems carry the same world-model and policy architecture and are trained to statistically indistinguishable competence on a common sensorimotor task suite before testing, with competence matching preregistered as an inclusion criterion for the comparison.

Perturbation battery, applied identically: masking-like degradation of input evidence across a range of signal-to-noise ratios; valence perturbations; selective ablation of world-model, policy, and metacognitive components; and confidence elicitation at matched objective performance.

Primary outcome: a preregistered manifold-similarity score quantifying how closely each system's pattern of dissociations matches the human manifold, computed by a party blind to which system produced which data. Preregistered subcomponents: whether valence perturbations reorganize attention, learning, memory, and action jointly or move a single score; whether confidence dissociates from accuracy in the human pattern; and whether ablations degrade a broad repertoire or only the benchmark built around the ablated component.

Falsification rule: System A scoring higher than System B on the manifold-similarity score, at matched competence, with the margin exceeding a preregistered threshold. §12 states the consequence: because the manifold is evidential rather than constitutive, this does not strictly prove the Indexing Thesis false, but it removes its only route to evidential support in the domain the thesis was built for, and the theory should then be abandoned rather than defended.

Organoid extension. Add two Brainoware-derived systems matched for tissue lineage, size, interface, training history, and task performance. System C receives fixed external perfusion and has no access to oxygen, glucose, pH, temperature, or waste variables except through the experimenter; these variables do not enter its policy. System D receives the same exteroceptive stream but also senses those viability variables and can regulate perfusion, nutrient delivery, and protective action through the same recurrent policy that performs the task. Emergency support is permitted but logged; routine maintenance must be endogenous. Before testing, preregister whether the organoid-computer assembly rather than its replaceable controller is the identity-bearing token.

Prediction: System C may show plasticity-dependent learning and high integration yet fail the cross-domain valence and ablation signatures; System D should more closely reproduce the human dissociation manifold if organoid-level indexing has been achieved. A result in which C robustly outranks D at matched competence and measured integration counts against IFT; a result in which D outranks C while integration is matched or lower counts against pure

causal-structure accounts. Because current organoid systems do not meet these conditions, the extension is a build specification rather than a claim about existing Brainoware.

S6.5 Prospective test for T4 (currently unrunnable; not a protocol)

Stated for completeness. The system required is a linguistically competent artifact satisfying the indexing condition (for which System B of F4 is the nearest specified design) trained from scratch on a corpus from which philosophy of mind, phenomenology, and first-person description of experience have been removed, with removal verified by a held-out classifier and by adversarial probing for paraphrased content. The prediction is that such a system will spontaneously generate reports that something is present which it cannot specify, and that its states have a character its own descriptions omit, in the absence of any training signal for such reports. Absence of those reports would count against T4, weakly, since a conscious artifact might lack metacognitive access to its own format. Until such a system exists, T4 is a bet and is scored as one.

S6.6 Counting subjects (T2 against exclusion)

§12 pairs complete callosotomy with craniopagus twins sharing a thalamic bridge as a two-by-two natural experiment in which IFT and IIT's exclusion postulate count differently (S3.17). The criteria below are what a count needs.

Two agents rather than one: (i) two independently reporting channels, each able to report stimuli the other cannot access and to report its own states, with preferences that can be elicited separately and can conflict; (ii) separately manipulable viability conditions, so that a homeostatic challenge (thirst, glucose depletion, hypercapnia within clinical limits) can be delivered to one candidate and not the other and produces valenced behavior, autonomic response, and learning only in the challenged candidate. Both criteria are required, because (i) alone is satisfied by a division of access within one subject, which is IFT's reading of the split brain, and (ii) alone is satisfied by two bodies that nobody doubts are two persons.

Split brain: the prediction is that (i) is at most partly satisfied and (ii) cannot be, because the hemispheres share one body, one internal milieu, and one set of viability conditions; a homeostatic challenge reaches both, and the valenced response is one organism's. IFT would be refuted by a patient in whom the two hemispheres respond to one and the same bodily state with opposite valence (one seeking, one avoiding, each in the channel it controls) and in whom a challenge can be made to matter to one hemisphere and not the other. Pinto et al. (2017) and the network results of Santander et al. (2025) describe divided access with one responding agent, which is the predicted pattern.

Craniopagus twins with a shared bridge: the prediction is that (ii) is satisfied, since each twin has her own viability conditions, and that (i) is satisfied in the ordinary way, so that the count is two whatever content crosses the bridge. IFT would be refuted by a pair in whom one and the same homeostatic challenge, delivered to one body, produced valenced behavior in both

as one agent, with no separable preference or report. The case is reported rather than studied (Dominus, 2011), and the criteria are stated so that it, or the next case, can be.