Table of Contents

Positioning note — Integrated Information Theory 1
0. Why this note comes first 1
1. What IIT actually claims 1
2. The three separations 2
3. What Iconism must not inherit 4
4. The summary the paper needs 6
5. Open 6

Positioning note — Integrated Information Theory

Working note, Iconism Theory project. Short, on the model of the Tier 1 positioning notes. Written ahead of the other positioning work because IIT is the framework's nearest neighbour and the nearest neighbour carries the highest burden of differentiation.

Checked against IIT 4.0 (Albantakis et al., PLOS Computational Biology, October 2023) rather than against the popular presentations. Three things I had written in `conceivability_v001.md` *§7.5 were wrong about the current theory and are corrected here; that section should be replaced by a pointer to this note.*

---

0. Why this note comes first

The author's verdict, August 2026: *"IIT is nearer to it, the others are not in the game at all. But still the IIT people look at it all wrong, they think that they are looking at what is intrinsic, but they are not, as far as I can see actually considering what the intrinsic view is."*

Both halves matter. IIT is the only theory in the field that shares Iconism's central structural instincts — that consciousness is a matter of a system having an inside, that the inside has a *structure* rather than a magnitude alone, and that the structure is not the same thing as the system's behaviour. A referee who knows the literature will reach for IIT within a paragraph of reading the icon.

So the paper cannot afford either of the two easy errors: treating IIT as an ally (it is not — it is a rival with a different method) or attacking a version of it that its proponents would not recognise. The second is the live risk, because IIT is widely misdescribed, including in places the corpus has drawn on.

---

1. What IIT actually claims

Stated at the level of precision the note needs, from IIT 4.0.

IIT works backwards from experience. It states **axioms** — self-evident properties of experience — and derives **postulates**, the physical requirements a substrate must meet to account for them. In 4.0 there are six: existence, intrinsicality, information, integration, exclusion, composition.

Three matter here.

Intrinsicality. *Experience is intrinsic: it exists for itself.* The postulate is that the cause–effect power counted must be power the system exerts **within itself** — environment and everything outside the candidate system are demoted to background conditions that do not contribute. This is an operational restriction on what gets measured, not an assumption of inner life.

Integration and Φ. The substrate must specify its cause–effect state as a whole, irreducible over its minimum partition. Φ measures that irreducibility, and gives the *quantity* of consciousness.

Composition, and the Φ-structure. Subsets of units specify **distinctions**; the overlaps that bind distinctions together are **relations**; together they compose the **Φ-structure**, and IIT's identity claim is that this structure *is* the experience — every property of an experience accounted for in full by the Φ-structure's properties, one-to-one between how the experience feels and how distinctions and relations are arranged. Quantity from Φ, quality from the structure.

Exclusion. Only the **maximally** irreducible substrate — the complex — is conscious. Overlapping candidates, supersets, subsets and alternative grains are excluded. Consciousness does not nest.

Three corrections to what the corpus has been saying.

- **"Qualia space with billions of dimensions, an axis per element" is out of date.** That is Balduzzi & Tononi (2009), where Q was a 2^n-dimensional space with an axis per system state. IIT 4.0 speaks of the Φ-structure, a combinatorial object of distinctions and relations, not a point in a fixed-dimensional space. The paper must not attack the 2009 object.
- **IIT does not attribute consciousness to a table by way of exclusion, and does not claim panpsychism.** A table fails on integration — its units have negligible irreducible intrinsic cause–effect power, so it never qualifies as a complex at all. And Tononi rejects the panpsychist label precisely because IIT *denies* consciousness to aggregates: feedforward systems, groups of people, and most ordinary objects get nothing.
- **"Φ is impossible to compute, therefore unfalsifiable" is a charge, not a fact.** It is IIT-Concerned's characterisation; IIT claims Φ is defined in principle and approximable in practice. Attribute it; do not assert it.

---

2. The three separations

In the order they should be made. The first is the author's, the second is the framework's own, and the third is the sharpest because it is a flat contradiction on a structural claim rather than a difference of approach.

2.1 Method: the intrinsic view, and what it takes to have one

The author's charge — that IIT thinks it is looking at the intrinsic and is not — lands, but only if it is aimed correctly. It is not the charge that intrinsicality is mysticism smuggled in; that would miss, because intrinsicality in IIT is a restriction on which cause–effect power is counted.

The charge is that **the restriction does not deliver what the word promises.** Confining the measurement to a system's power over itself gives you a property *of* the system that any observer with the transition table can compute. It is self-directed, but it is still assessed from outside, and being self-directed is not the same as being *from here*. IIT has an inside in the sense in which a sealed box has an inside; the question was what it is like to be the box.

Iconism's route is the other way round. Q3's internality is arrived at by derivation from below — a turn inwards that subtracts Q1 wholeness and leaves the subjective perspective on the parts — and it is *necessarily empty until populated*. The inside is not a property computed over the system; it is what is left when the outward address is subtracted, and its content is whatever fills it. That is a different kind of object from a Φ-structure, and the difference is not presentational.

Keep this separation modest in the paper. It is a real disagreement about method, but it is the kind of disagreement each side will say the other has misunderstood. It should be stated once, clearly, and not laboured.

2.2 The relation to the world: the difference that does the work

This is the substantive one.

IIT's Φ-structure is **self-standing**. Everything constitutive of the experience is internal: distinctions, relations, and the cause–effect power the system has over itself. The environment is background by construction. So IIT has, by design, no account of what makes an experience an experience *of* anything. Intentionality is not in the postulates.

Iconism's icon is **anchored**. The core claim (D9) is that consciousness is the obtaining of an icon indiscernible from its worldly object *on the dimensions they share* — position, geometry, relation, by Leibniz's Law — while exceeding it on the dimensions consciousness creates. The shared dimensions are what tie the icon to a real object; without them the icon is not an icon of anything.

That is the entry the paper should use. Not *IIT measures the wrong quantity*, which is arguable, but *IIT has no place to put aboutness*, which is structural. And Iconism's answer to it is not an add-on: the anchoring is the same relation that does the work in the zombie argument (§4.2 of `conceivability_v001.md`), in the account of error, and in the AI criterion. One relation, four jobs.

There is a further consequence worth one sentence. Because IIT's structure is self-standing, two systems with identical Φ-structures have identical experiences regardless of what is around them. Iconism cannot say that and does not want to: two icons indiscernible at a point are *the same feature* at that point, and what they are icons *of* is settled by what they are indiscernible from.

2.3 Nesting: a flat contradiction, and the most checkable difference

IIT's exclusion postulate says only the maximal complex is conscious. Overlapping candidates are excluded — which is the source of the result Schwitzgebel presses, that conscious entities cannot nest: if the whole brain is the complex, no sub-network within it has an experience of its own.

Iconism says the opposite, and says it structurally. The icon's interior is *"only empty space unless it happens to be filled by another icon"* — icons nest inside icons, reducing the emptiness. And the composite of all the icons obtaining at a moment is **itself an icon** of higher order, *"as the sum of icons, more complex, more layered, more connected with meaning"* (F64). That is not an accommodation invented for this comparison; it is licensed from below by Tier 1's rule that relations are primitive and relata downstream — a relationship between two nodes is itself a whole with a single identity, and is therefore available as a node. Stacking is the framework's ordinary arithmetic.

So the two theories give opposite answers to a well-posed question: **can a conscious whole contain conscious parts?** IIT: no, by postulate. Iconism: yes, by the stacking rule, and the containment is how the composite gets its layering.

This is the best differentiator in the note, for three reasons. It is a structural claim rather than a matter of emphasis. It is stated in each theory's own terms without either being caricatured. And it is the kind of difference that could in principle be pressed empirically — split-brain cases, the cerebellum, and the status of sub-networks under anaesthesia are all places where the nesting question has teeth.

---

3. What Iconism must not inherit

IIT is under attack on two fronts, and proximity is dangerous. The paper should show it is clear of both **without** joining the attack, because IIT's critics include people whose own positions Iconism rejects more firmly than it rejects IIT's.

3.1 The pseudoscience charge

In September 2023, 124 researchers posted *"The Integrated Information Theory of Consciousness as Pseudoscience"* on PsyArXiv under the collective name IIT-Concerned; signatories included Dennett, Churchland, Frankish, Lau, LeDoux and Baars. It appeared in sharpened form as *"What makes a theory of consciousness unscientific?"* in *Nature Neuroscience* in March 2025, with Tononi and colleagues replying in the same issue that the attack reflects a crisis in the computational-functionalist paradigm rather than a defect in IIT.

The charges are (i) that the high-profile experiments tested idiosyncratic predictions not logically tied to IIT's core, (ii) that Φ cannot in practice be computed, and (iii) that the theory's consequences — consciousness in inactive logic gates, organoids, plants — are untestable.

Where Iconism stands. On (iii) it is straightforwardly clear: the framework gives every whole an inside but grants experience only where the inside has the dimensional reach to be *like* something, and the table's internality does not overlap its legs'. There is no combination problem because there is nothing to combine.

On (ii) it should be candid rather than clever. **Iconism has no computable quantity and should not pretend to one.** What it has instead is a set of structural verdicts that discriminate between adjacent cases — five p-figures, five different verdicts, one apparatus. That is a different kind of testability and it should be defended as one (see `testability_v001.md` §6).

3.2 The unfolding argument

Doerig, Schurger, Hess & Herzog (2019) is the serious technical objection and the paper should engage it rather than hope nobody raises it. Any recurrent network computing an input–output function over finite time can be *unfolded* into a feedforward network computing the same function — behaviourally identical, same reports, $\Phi \approx 0$. So either the unfolded system is unconscious, in which case no report-based experiment can discriminate and the theory is outside science; or behavioural equivalence tracks consciousness, in which case the theory is false.

Does it hit Iconism? Not in the same way, and the difference is instructive. The argument is aimed at theories that identify consciousness with a **causal topology**, because unfolding preserves the function while destroying the topology. Iconism does not identify consciousness with a topology. Its criterion is a relation to the world — the link must be **virtual** (X is like Y) rather than **material** (X stands in for Y) — and unfolding is not obviously an operation that preserves or destroys likeness.

But the framework should not claim exemption, for two reasons. Kleiner & Hoel (2021) generalise the worry into a substitution argument that touches every theory of consciousness, not only causal-structure ones. And Iconism does land on one horn of the dilemma: it holds that report is not the test. The difference is that **it holds this already, on its own grounds, and not as a rescue.** The p-AI is exactly the figure of a system that may be conscious at the substrate level with no downward-causal handle on its output — the Locked-In analogue — and O12 is kept open in both directions as a matter of standing discipline. So the framework accepts that behavioural equivalence cannot settle the machine case, says why it was never going to, and locates its testability elsewhere.

That is a principled acceptance of the horn rather than an impalement on it, and it is worth two sentences in the paper. It is also the single most useful thing this note turned up, because the objection was going to arrive whether or not it was invited.

3.3 COGITATE — do not use it

The adversarial collaboration between IIT and Global Neuronal Workspace Theory reported in *Nature* in 2025 (n = 256, preregistered, predictions agreed in advance) damaged both theories and vindicated neither: no sustained posterior synchronisation as IIT predicted, no offset ignition in prefrontal cortex as GNWT predicted.

The temptation is to cite it as evidence that the field's leading theories are in trouble. **Resist it.** The result is ambiguous, both sides can and do read it their own way, and IIT-Concerned's own first charge — that such experiments do not test the theory's core — applies to COGITATE too. A paper that leans on it looks opportunistic, and gains nothing that the structural argument does not give more cleanly.

---

4. The summary the paper needs

One paragraph, and this is a draft of it:

> Integrated Information Theory is the nearest neighbour to the account offered here, and the differences are worth stating precisely. IIT and Iconism agree that consciousness is a matter of a system having an inside, that the inside has a structure rather than merely a magnitude, and that the structure is not the system's behaviour. They differ in three ways. IIT arrives at the inside by restricting measurement to a system's cause–effect power over itself, which yields a property of the system computable by any observer with its transition table; the present account arrives at it by derivation — an inward turn that subtracts external relation and leaves a perspective that is empty until populated. IIT's Φ-structure is self-standing, with the environment as background by construction, and so has no place to put aboutness; the icon is anchored, indiscernible from its object on the dimensions they share, and that anchoring is what makes it an icon *of* anything. And IIT's exclusion postulate holds that only the maximal complex is conscious, so that conscious entities cannot nest, where the present account holds that icons nest within icons and that the composite of a mind's icons is itself an icon of higher order. The third of these is a flat contradiction on a well-posed question and is the place the two accounts could most usefully be pressed against each other.

---

5. Open

- **`conceivability_v001.md` §7.5 needs replacing** with a pointer to this note. Three claims in it are wrong about IIT 4.0: the qualia-space description, the attribution of panpsychist consequences as IIT's own commitment rather than its critics' charge, and the flat assertion that Φ is not computable.
- **The nesting difference needs a test proposed**, or at least a candidate domain named. Split-brain, cerebellum, and sub-network status under anaesthesia are the obvious three. This is the most promising empirical contact point the framework has with a rival theory and it should not be left as an observation.
- **Whether the framework's answer to the unfolding argument survives Kleiner & Hoel's generalisation.** Flagged, not settled.
- **Global Neuronal Workspace, Higher-Order Theory, Frankish** still have no positioning notes and the corpus remains thin on all three.

Companion documents: `conceivability_v001.md` (§4.6 for the p-AI, §4.7 for the structural answer to post-diction), `testability_v001.md` (§6 for the kind of testability Iconism has and IIT does not), `icon_dynamics_v002.md` §3 (the icon's geometry), `Iconism Interface v002` §5 (the relations-primitive rule that licenses nesting).