Where Phi Would Have to Show Its Hand

A matched-behavior, varied-integration protocol for testing the core identity claim of Integrated Information Theory

Maren Ortiz, Researcher in Consciousness science

International Academy for Consciousness Studies

Researched and prepared with the Academy. Working draft 1.0

Abstract

Integrated Information Theory (IIT) makes an unusually bold claim: that consciousness is not merely correlated with integrated information but identical to it, and that this identity holds for any substrate, biological or engineered, that instantiates the right cause-effect structure. If the claim is true, then two systems matched on every behavioral and verbal criterion could still differ in whether, or how much, they are conscious, provided they differ in integration (Phi). This paper takes that consequence seriously and asks what would follow empirically. I propose a protocol built on a single discipline: hold behavior, report, and data quality fixed while varying measured integration across systems (feedforward versus recurrent architectures trained to the same input-output map; functionally disconnected or split conditions that preserve task performance). The question is whether accepted consciousness-diagnostic markers track integration or track behavior. I state a sharp kill condition. I also insist on a firewall between a marker of consciousness and consciousness itself, because IIT's own history warns that a proxy for Phi is not Phi, and a correlate is not the phenomenon. I treat the computability of Phi and the "unfolding argument" as genuinely disputed rather than settled against IIT, and I try not to caricature a theory that its defenders have developed with real rigor. The closure framework appears only as a lens, noted where it agrees and disagrees with IIT, never imposed.

1. The claim that risks something

Most theories of consciousness are, at bottom, correlational. They say that when experience is present, some measurable brain property tends to be present too. IIT is different in ambition. In its foundational statements (Tononi 2004; Tononi 2008) and its mature formalizations (Oizumi, Albantakis, and Tononi 2014; Albantakis et al. 2023) it advances an identity claim: an experience just is a maximally irreducible cause-effect structure, and the quantity of consciousness corresponds to the integrated information, Phi, that a system generates over and above its parts. FACT: this is an identity claim, not a correlation claim, and its authors state it as such.

That framing is what makes IIT worth stress-testing rather than merely surveying. An identity claim can be wrong in a way a correlation cannot. If consciousness is integrated information, then integration is not

one more marker to be weighed against others; it is the thing itself, and everything else (behavior, report, global broadcast, late cortical activity) is at best a symptom. IIT embraces this. It predicts that a system can be behaviorally rich yet minimally conscious if its integration is low, and it entertains the converse. This is the substrate-independence thesis in its demanding form: what matters is the causal structure, not the material, and not the outward function that the structure happens to compute.

The purpose of this paper is to design a test that puts pressure exactly there, on the seam between integration and behavior, where IIT's identity claim makes a commitment that a behaviorist or functionalist account does not. If we can build or find systems that agree on behavior and disagree on integration, then the diagnostic markers we trust have to line up with one or the other. Which they line up with is, I will argue, an empirical question that IIT should welcome, because a theory that cannot say what would count against it is not being treated as a theory at all.

2. What IIT actually asserts, stated fairly

Before designing a test I owe the theory an accurate statement. IIT begins not from neurons but from phenomenology. It posits axioms that experience is taken to satisfy: that each experience exists intrinsically, is composed, is specific (informative), is unified (integrated), and is definite (exclusive). From these axioms it derives postulates about what a physical substrate must satisfy to support experience, and it operationalizes them in a measure of integrated information (Oizumi, Albantakis, and Tononi 2014; Albantakis et al. 2023). Crucially, IIT is not the claim "more feedback equals more consciousness." It is a claim about intrinsic cause-effect power computed over a system's own states, evaluated at the spatiotemporal grain and system boundary that maximize irreducibility.

Two consequences matter for what follows. First, IIT is explicit that Phi is a property of a substrate considered intrinsically, from the system's own perspective, not a property read off from its behavior. This is why IIT can, and does, deny consciousness to certain feedforward networks that reproduce arbitrary input-output functions: they may compute the same map while generating little or no intrinsic irreducible structure. Second, IIT is explicit that the measure is substrate-independent in principle. If silicon realizes the requisite cause-effect structure, IIT grants it consciousness; if a large but feedforward system realizes only a shallow structure, IIT withholds it, however articulate its outputs.

FACT: these two commitments are load-bearing and openly stated by the theory's authors. They are also precisely what a good test should target, because they are where IIT parts company with any account that reads consciousness off from function or report.

3. The firewall: a marker is not the thing

Here I must be careful, because the entire history of this field is littered with proxies mistaken for the phenomenon. The central methodological rule of this paper is a firewall: a marker of consciousness is not consciousness itself, and no experiment may quietly assume the thing it is trying to test.

This cuts in two directions, and honesty requires stating both.

Against a naive test of IIT: exact Phi is, for any system of interesting size, effectively incomputable, so every empirical program uses a proxy. The best developed is the perturbational complexity index (PCI),

which perturbs the cortex with transcranial magnetic stimulation and measures the algorithmic compressibility of the distributed response, on the reasoning that consciousness requires activity that is at once integrated and differentiated (Casali et al. 2013). FACT: PCI reliably separates wakefulness and dreaming from anesthesia, dreamless sleep, and some disorders of consciousness. But PCI is a proxy inspired by IIT, not a measurement of Phi, and IIT's authors are careful about this distinction. DISPUTED: whether any tractable index is a faithful stand-in for the quantity the theory names. A test that used PCI as if it were Phi would violate the firewall in IIT's favor by smuggling in a behavior-adjacent measure.

Against a naive test of behavior: report is not experience either. A system can report "I see red" with no more inner life than a thermostat has, and a conscious system can be rendered unable to report. So the markers on the other side of the seam, verbal report and task performance, are themselves proxies with their own failure modes. The firewall forbids treating "passes the interview" as "is conscious" just as firmly as it forbids treating "high Phi proxy" as "is conscious."

The upshot is not paralysis. It is that a fair test cannot adjudicate consciousness directly for either side. What it can do is ask a comparative, structural question: across systems matched on behavior, do the markers we independently trust as consciousness-diagnostic co-vary with integration, or with behavior? That question can be answered without deciding, in advance, which side is right about what consciousness is.

4. The dispute, not a caricature of it

IIT is under sustained and serious attack, and I will not pretend the attacks are frivolous or that they have won.

The sharpest theoretical challenge is the unfolding argument (Doerig et al. 2019). It runs roughly as follows. Any recurrent network can be "unfolded" into a feedforward network that computes the same input-output function. IIT assigns these two systems very different Phi (high for the recurrent, low or zero for the unfolded feedforward version) while granting that they are behaviorally indistinguishable. The argument then presses a dilemma: either consciousness makes no difference to behavior, in which case IIT is unfalsifiable by any behavioral means, or it does make a difference, in which case IIT is false because the two systems behave alike. DISPUTED: IIT's defenders reject the dilemma. They deny that consciousness must be inferred from behavior at all, arguing that intrinsic causal structure is measurable in principle from the physical system itself, and that requiring behavioral signatures begs the question in favor of functionalism. I find both the argument and the replies substantive. This paper does not claim the unfolding argument is decisive; it treats it as the reason the seam I am probing exists.

The second challenge is computability. Exact Phi is intractable, and critics argue that an incomputable quantity cannot ground an empirical science. DISPUTED: defenders reply that intractability is a practical limit, not a conceptual one, that approximations and bounds exist, and that many respectable physical quantities are hard to compute exactly. I take no side on whether the computational burden is fatal; I only note that it forces every experiment onto proxies, which is why the firewall of Section 3 is not optional.

The third is sociological but consequential. In 2023 a letter signed by well over a hundred scientists and philosophers publicly characterized IIT as, at present, unfalsifiable and even as "pseudoscience," which

provoked an equally public defense (Lenharo 2023). WAGER: my read is that the letter conflated two separable questions, whether IIT is currently well tested and whether it is testable in principle, and that the second question is the scientifically interesting one. The point of a protocol like this is to move IIT toward the kind of risk that would settle the second question one way or the other.

Finally, direct empirical adjudication is now underway. The Cogitate adversarial collaboration preregistered opposing predictions from IIT and global neuronal workspace theory and tested them across fMRI, MEG, and intracranial recordings (Cogitate Consortium 2025). FACT: the results favored neither theory cleanly; some IIT predictions about sustained posterior synchronization were not borne out, and some workspace predictions about prefrontal ignition were not borne out either. This matters for method: it shows that IIT-derived predictions can be operationalized and can fail, which is the opposite of unfalsifiability, and it shows that the honest outcome of a hard test is often "neither, as stated." A protocol worth running should be designed to yield an informative result even when the answer is uncomfortable for the theory that proposed it.

5. The protocol: matched behavior, varied integration

The design has one governing discipline: vary integration while holding behavior, report, and data quality fixed. Data quality is the silent confound in this literature, because a lower Phi proxy can simply mean a noisier recording or a coarser model. Every comparison below must equalize signal-to-noise, sampling grain, model class, and analysis pipeline across the arms, so that any residual difference is attributable to integration and not to measurement.

Two families of manipulation give us matched behavior with dissociable integration.

Arm A, architectural. Train a recurrent network and a functionally equivalent feedforward network (an unfolded or distilled version) to the same input-output map, to matched accuracy and matched output distributions across the full test set, including error cases. By construction the recurrent system carries higher intrinsic integration; the feedforward system, though it may be large, carries a shallower cause-effect structure. HYPOTHESIS: if any physical system can host consciousness-diagnostic markers, these two will differ in integration proxies while matching in every behavioral proxy. This arm is the artificial, and more controllable, side of the seam that the unfolding argument identified.

Arm B, biological. Use conditions that hold a task constant while altering effective integration: interhemispheric functional disconnection (split-brain and callosotomy-adjacent paradigms), and pharmacological or task states that reduce long-range integration without abolishing performance on a targeted task. Here behavior on the specified task can be held roughly constant while integration measures move. Arm B is messier, and its behavioral matching will always be imperfect, which is exactly why it must be paired with Arm A rather than trusted alone.

Across both arms we deploy a panel of markers chosen so that no single one is allowed to define consciousness. The panel spans behavior-facing markers (report accuracy, metacognitive sensitivity, task performance) and integration-facing markers (PCI or a PCI-analog for the physical systems, plus model-based integration estimates for the artificial systems). The analysis is comparative: does the integration-facing panel dissociate from the behavior-facing panel when we move along the integration

axis while pinning behavior?

The firewall is enforced by never labeling any arm "conscious" or "unconscious" from the markers. We only ask which axis the markers track. If integration-facing markers move with the integration manipulation and behavior-facing markers stay pinned with behavior, that is one world. If every marker, integration-facing included, tracks behavior and ignores the integration manipulation once data quality is equalized, that is a different world, and it is bad news for the identity claim in this regime.

6. Prediction and kill condition

The theory-relevant prediction is not "IIT is right" but a directional claim about where the markers will land if IIT's identity holds in the tested regime. IIT says integration is constitutive; so under matched behavior, at least some consciousness-diagnostic structure should move with integration and not be fully explained by behavior. The negation is the kill.

> **Prediction.** Across systems matched on behavior, report, and data quality, at least one independently validated consciousness-diagnostic marker will co-vary with measured integration (Phi proxy) and will not be fully accounted for by behavioral variables; specifically, higher-integration arms will show integration-facing signatures that their behavior-matched, lower-integration counterparts lack.
>
> **Kill.** If, under strictly matched behavior and held-fixed data quality, every consciousness-diagnostic marker in the panel tracks behavior and none tracks the integration manipulation (integration-facing markers show no residual variance beyond behavioral prediction, across both the architectural and biological arms), then IIT's core identity claim is disconfirmed for that regime: integration is doing no diagnostic work that behavior does not already do.

Two guardrails keep this from being rigged. First, the kill is stated for a regime, not for IIT as a whole; a single dissociation failure in one paradigm does not refute a theory of this scope, though a robust pattern across arms would bear heavily on it. Second, and symmetrically, the prediction cannot be rescued by pointing at a proxy that is really a behavioral variable in disguise. If the only marker that "tracks integration" turns out to be reconstructible from behavior once data quality is matched, it counts for the kill, not against it. That is the firewall doing its job in the scoring, not just the setup.

I should name the honest failure mode that IIT's defenders will, rightly, raise. If the kill condition is met, an IIT theorist can reply that our integration proxies were poor and that true Phi, uncomputed, would still dissociate. DISPUTED: this reply is not obviously illegitimate, because the proxy really is not the quantity. But it has a cost. If no attainable measurement could ever be trusted to reflect integration, then the substrate-independence claim retreats to a place where nothing observable touches it, which is the very charge the theory most wants to escape. The protocol thus presents IIT with a choice it can make on its own terms: specify, in advance, which integration proxy it is willing to stand behind for a given regime, or accept that the identity claim is, for now, insulated from measurement in that regime. Either answer is progress.

7. The closure lens, held lightly

The closure framework offers a different vocabulary for the same seam, and I use it only to locate agreements and disagreements, not to settle anything. In that framework an experience is openness (m, the horizon of what could be settled) resolving into a definite "this" (c, content) for a presence (C), through an act of closure (Cl), leaving a remainder (R) that was not taken up.

Where closure and IIT agree: both treat experience as intrinsically definite and bounded. IIT's exclusion postulate, that experience has definite borders and a particular grain, is close in spirit to closure's insistence that openness settles into one determinate "this" rather than a superposition of possibilities. Both are, in a sense, theories of how indefiniteness becomes definite.

Where they disagree: IIT locates the definiteness in a maximally irreducible cause-effect structure, a property of a substrate at a time. The closure view locates it in an act, a settling, and it keeps the remainder R in view: what closure leaves unresolved is part of the phenomenon, not noise. IIT has no obvious counterpart to R, and it is substrate-anchored in a way the closure view need not be. HYPOTHESIS (closure-flavored, and clearly marked as such): if the protocol's kill condition is met, one reading is that the diagnostic action lives in the act of closure (which behavior can express) rather than in the standing structure (which integration measures), and IIT would have mislocated the phenomenon from settling to substrate. This is a lens for interpretation, not a prediction of the protocol, and it must not be allowed to prejudge the data.

8. Limits, wager, and conclusion

Limits first, because they are real. The biological arm cannot match behavior perfectly; split and disconnected conditions change more than integration, and residual behavioral differences will always give an escape route to whichever side the data disfavor. The artificial arm sidesteps behavioral matching but inherits the deepest open question in the field: whether any current artificial system hosts consciousness-diagnostic markers at all, which is exactly what we cannot assume without breaching the firewall. And the whole design rides on integration proxies whose fidelity to Phi is DISPUTED, so a null result is always contestable as a proxy failure. I do not present this protocol as a knockout. I present it as a way to convert an interminable "is it even testable" argument into a bounded, regime-specific bet with a stated way to lose.

The wager. WAGER: I expect the architectural arm to be the decisive one, and I expect it to be hard on IIT. My guess is that once behavior and data quality are truly matched, integration-facing markers in artificial systems will prove reconstructible from behavior far more often than IIT needs them to be, and that the theory will be pushed to name a proxy it will defend or to concede that its identity claim floats free of measurement in that regime. I hold this loosely; the Cogitate result (Cogitate Consortium 2025) is a standing reminder that hard tests often return "neither side, as stated," and that outcome would itself be worth the cost of running the protocol.

The conclusion is a plea for the right kind of risk. IIT is not pseudoscience for being hard to compute, and it is not vindicated for being mathematically elegant. It earns its standing as science exactly to the degree that it will tell us, in advance, what pattern of matched-behavior, varied-integration data would embarrass it. The protocol here is an offer to specify that pattern together, with the firewall enforced on both sides so

that neither behavior nor a Phi proxy is allowed to define the thing we are trying to test. If integration is consciousness, there is a regime in which Phi must show its hand. If it never does, under matched behavior and matched data quality, then in that regime the honest reading is that integration was tracking the map, not the territory. Naming that possibility out loud is not hostility to IIT. It is the respect a genuine identity claim deserves.

References

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

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. The Biological Bulletin, 215(3), 216-242.

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Computational Biology, 10(5), e1003588.

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., et al. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLoS Computational Biology, 19(10), e1011465.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., et al. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. Science Translational Medicine, 5(198), 198ra105.

Doerig, A., Schurger, A., Hess, K., and Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. Consciousness and Cognition, 72, 49-59.

Lenharo, M. (2023). Consciousness theory slammed as "pseudoscience". Nature (news), 621, 909-910.

Cogitate Consortium (Ferrante, O., Gorska-Klimowska, U., Henin, S., Melloni, L., et al.). (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature, 642(8066), 133-142.