Toward Operational Terminology in Integrated Information Theory: A Methodological Consideration

Exploring Alternative Conceptual Frameworks for IIT 4.0

Boris Kriger

Institute of Integrative and Interdisciplinary Research
boriskriger@interdisciplinary-institute.org

Abstract

This paper explores the possibility of supplementing the conceptual vocabulary of Integrated Information Theory (IIT 4.0; Albantakis et al., 2023) with operationally defined terms when discussing its mathematically rigorous constructs. Without critiquing IIT's formalism or denying that Φ-structures may correspond to phenomenal experience, we suggest that certain methodological advantages might be gained by distinguishing between the theory's quantifiable measures (cause-effect repertoires, integrated information Φ, distinctions, relations, maximal substrates) and the interpretive claim that these measures are identical to consciousness. We tentatively propose that in strictly empirical and computational contexts, supplementing IIT's vocabulary with operationally neutral terms—free of centuries-old phenomenological and metaphysical baggage—may help avoid framing effects, illusory consensus, and interdisciplinary confusion. The specific labels we use are illustrative only; what matters is the systematic separation of what IIT rigorously quantifies from the age-old interpretive leap to subjective experience.

Keywords: integrated information theory, IIT 4.0, operational definitions, methodology, Φ-structure, cause-effect structure

1 Introduction

1.1 Background and Motivation

Integrated Information Theory (IIT), particularly in its recent 4.0 formulation (Albantakis et al., 2023), represents one of the most systematic and mathematically rigorous attempts to characterize consciousness in physical terms. The theory develops an impressive formal apparatus: transition probability matrices, cause-effect states, intrinsic information measures, integration quantified through partitioning, and the unfolding of Φ-structures with distinctions and relations. This mathematical machinery is precise, computable (at least for small systems), and generates specific, testable structural predictions.

At the same time, IIT 4.0 makes strong identity claims connecting this formalism to phenomenal experience. The authors state that "an experience is identical to the Φ-structure" and that "the Φ-structure corresponds to the quality of the experience" (Albantakis et al., 2023). They formulate "the properties of phenomenal existence in physical terms" and ground their postulates in phenomenological axioms described as "immediate and irrefutable."

The present paper does not question these identity claims. Rather, we explore a narrow methodological question: when discussing IIT's quantifiable measures in empirical and computational contexts, might there be pragmatic advantages to using terminology that does not immediately invoke the full weight of philosophical debates about consciousness?

1.2 Scope, Limitations, and Disclaimer

It is essential to state clearly what this paper does and does not propose.

This paper is a purely methodological illustration. We do NOT propose abandoning the concept of consciousness in everyday, clinical, ethical, or philosophical discourse. We do NOT critique the mathematical formalism of IIT 4.0 (Albantakis et al., 2023)—we consider it an impressive achievement. We do NOT deny that Φ-structures may correspond to phenomenal experience. We take no position on whether the identity claim at the heart of IIT is true or false.

Our sole aim is to demonstrate one pragmatic approach: when discussing the rigorously quantifiable aspects of IIT (cause-effect repertoires, intrinsic information, integrated information Φ, distinctions, relations, maximal substrates), it may be useful to employ neutral, operational terminology free of centuries-old phenomenological and metaphysical baggage. This modest terminological hygiene may help avoid framing effects, illusory consensus, and interdisciplinary confusion—without touching the theory's core identity claim.

The specific terms we propose are merely illustrative. Any neutral vocabulary that separates the mathematical formalism from its phenomenological interpretation would serve the same purpose.

1.3 The Concern: Brilliant Formalism, Medieval Debates

Our motivation can be stated simply: it seems regrettable when years of rigorous mathematical work become entangled in debates that more closely resemble medieval scholasticism than contemporary science.

IIT 4.0 develops sophisticated tools for characterizing the intrinsic causal structure of physical systems. These tools—Φ calculations, cause-effect state analysis, distinction and relation identification, maximal substrate determination—are mathematically well-defined and, in principle, empirically applicable. This is genuine scientific progress.

Yet the moment these tools are described as measuring "consciousness" or "phenomenal experience," a different set of debates is activated: Does a thermostat have micro-consciousness? What about a lookup table with high Φ? Are we all zombies? Is this just panpsychism in mathematical dress? These are legitimate philosophical questions, but they risk overshadowing the scientific contributions of the formalism itself.

Our suggestion is simple: distinguish more carefully between (a) what IIT's mathematics demonstrably characterizes—integrated cause-effect structures—and (b) the further claim that such structures are identical to phenomenal experience. One can accept (a) while remaining agnostic about (b). And for researchers primarily interested in (a), having vocabulary that does not presuppose (b) might prove useful.

1.4 Structure of the Paper

Section 2 discusses the general challenge of historically laden terminology in science. Section 3 proposes some illustrative operational concepts. Section 4 considers how these relate to IIT 4.0's framework. Section 5 discusses potential applications, and Section 6 addresses limitations and objections.

2 Terminological Considerations

2.1 Broad Historical Concepts and Their Scientific Limitations

Before addressing consciousness specifically, it may be useful to situate this discussion within a broader methodological context. The sciences—both natural and social—have inherited a num-

ber of broad, historically laden concepts that, while culturally significant, present considerable challenges for rigorous inquiry.

Consider, for example:

- **Soul**: Perhaps the paradigmatic case of a concept that once played a central explanatory role but has largely been set aside in scientific contexts. While "soul" remains meaningful in theological and everyday discourse, the phenomena it was invoked to explain—life, thought, personal identity, moral agency—are now typically addressed through more specific biological, psychological, and neuroscientific constructs. This transition did not require proving that souls do not exist; it simply reflected a methodological preference for more operationally tractable concepts.

- **Freedom**: A term central to political philosophy, ethics, and everyday discourse, yet one that encompasses incompatible meanings (negative liberty, positive liberty, free will, political autonomy, freedom of choice) and carries substantial ideological baggage. Empirical research on decision-making, behavioral economics, or political systems often proceeds more productively when specific, operationally defined constructs replace the undifferentiated concept of "freedom."

- **Justice**: Another foundational concept with millennia of philosophical elaboration, yet one that means different things in different contexts (distributive justice, procedural justice, restorative justice, retributive justice). Legal and social scientific research frequently benefits from specifying which operationally defined aspect of "justice" is under investigation.

- **Conscience**: A concept with deep theological and philosophical roots, often invoked in moral discourse but rarely defined with precision. Research on moral reasoning, ethical decision-making, or prosocial behavior typically employs more specific constructs rather than the diffuse notion of "conscience."

- **Consciousness**: The focus of the present paper, similarly characterized by multiple incompatible definitions and substantial philosophical baggage.

What these concepts share is a combination of (a) intuitive resonance—they name something that seems important and real, (b) definitional ambiguity—they are used in multiple, not always compatible ways, and (c) historical accumulation—centuries of usage have layered meanings, associations, and connotations that resist precise specification.

2.2 Potential Difficulties Arising from Inherited Terminology

It may be worth considering some of the difficulties that can arise when scientific inquiry employs terms inherited from pre-scientific or philosophical traditions without sufficient critical examination.

Illusory agreement: When researchers use the same term while meaning different things, apparent consensus may mask substantive disagreement. Two scientists claiming to study "consciousness" may in fact be investigating quite different phenomena, leading to confusion when their findings are compared or integrated. IIT 4.0's precise mathematical definitions could help resolve such confusion—but only if the terminology used to describe the mathematics does not reintroduce ambiguity.

Hidden assumptions: Historical terms often carry implicit metaphysical commitments that may go unnoticed. The term "consciousness," for instance, may implicitly suggest a unified, bounded entity—an assumption that IIT examines rigorously through its exclusion postulate but that casual terminology may take for granted.

Framing effects: The choice of terminology can shape how problems are conceptualized. Framing research as investigating "consciousness" may predispose researchers toward certain

questions (e.g., "Is this system conscious or not?") while obscuring others (e.g., "What is this system's integrated cause-effect structure?"). The latter question is mathematically precise; the former invites philosophical debate.

Communication barriers: Terms with heavy philosophical baggage can impede communication between disciplines. A neuroscientist, a philosopher, a computer scientist, and a clinician may all use "consciousness" while operating with substantially different assumptions about what the term denotes. IIT's formal precision could bridge these gaps—if the terminology does not reactivate the very confusions the formalism was designed to avoid.

Resistance to engagement: Some researchers may dismiss IIT's mathematical contributions simply because they are presented in the language of "consciousness studies." A presentation emphasizing the formal characterization of intrinsic causal structure might reach audiences who would otherwise disengage.

We do not claim that these difficulties are insurmountable or that inherited terminology should be abandoned wholesale. However, awareness of these potential pitfalls may encourage greater care in distinguishing between what IIT's mathematics demonstrably characterizes and how that characterization is interpreted.

2.3 The Specific Case of "Consciousness" in IIT 4.0

IIT 4.0 (Albantakis et al., 2023) is explicit about its phenomenological starting point. The authors begin with "essential properties of every experience" described as "immediate and irrefutable," including that experience "exists," is "structured," is "specific," is "unitary," and is "definite." These phenomenological axioms are then translated into physical postulates (existence, intrinsicality, information, integration, exclusion) that a system must satisfy to be a "substrate of consciousness."

The theory then develops precise mathematical tools for evaluating whether and to what extent a system satisfies these postulates. The culmination is the Φ-structure: "The Φ-structure is the quality of experience—what the experience is" and "the sum of $\varphi_{d/r}$ values corresponds to its quantity (Φ)" (Albantakis et al., 2023).

This is a bold theoretical move: identifying a mathematically defined structure with phenomenal experience. Our concern is not whether this identification is correct—that is a question we do not address—but whether the terminology used to discuss the mathematical structure need always invoke the identification.

Consider the analogy: physicists can discuss the mathematics of general relativity without constantly invoking the philosophical debates about the nature of time, substantivalism versus relationalism, or the reality of spacetime points. The mathematics stands on its own; philosophical interpretation is a separate (and valuable) enterprise.

Similarly, one might discuss IIT's Φ-structure analysis, its methods for identifying maximal substrates, and its techniques for unfolding cause-effect structures—all without constantly invoking debates about qualia, the hard problem, or panpsychism. For some purposes, this separation might prove useful.

2.4 A Historical Observation

It may be worth noting—without drawing overly strong conclusions—that the history of science includes cases where explanatory concepts were gradually supplemented or refined by more specific terminology. The concept of "life," for instance, was once thought to require a special vital principle; today, biological phenomena are typically explained in terms of more specific mechanisms (metabolism, reproduction, homeostasis, etc.) without necessarily denying that organisms are "alive."

This does not mean that "consciousness" will or should follow the same trajectory. But it suggests that supplementing a general concept with more specific operational terms is not

inherently problematic and may sometimes prove useful.

3 Illustrative Operational Concepts

In this section, we tentatively propose several concepts that might supplement existing terminology when discussing IIT's formalism in strictly empirical or computational contexts. These proposals are offered in an exploratory spirit and are explicitly illustrative. We do not claim that these specific terms are necessary—any neutral vocabulary that separates the mathematical characterization from the phenomenological interpretation would serve the same purpose.

3.1 Integrated Cause-Effect Structure (ICES)

Tentative definition: A system exhibits Integrated Cause-Effect Structure to the extent that it satisfies the postulates specified by IIT 4.0—namely, that it has intrinsic cause-effect power that is specific, integrated, and definite.

Relation to IIT formalism: ICES corresponds directly to what IIT 4.0 characterizes through its Φ-structure analysis. The term simply provides a way to refer to IIT's mathematical measures without immediately invoking the claim that these measures are identical to phenomenal experience. Albantakis et al. (2023) state that the postulates "are necessary and sufficient for a system to be a substrate of consciousness." ICES terminology allows one to discuss whether a system satisfies the postulates without presupposing what satisfying them implies about experience.

Note: This is merely one possible term. "Maximal irreducible causal structure," "intrinsic causal specificity," or simply "Φ-structure" (used descriptively) would serve equally well.

3.2 Internal Correspondence (IC)

Tentative definition: A system exhibits Internal Correspondence to the extent that its outputs (behaviors, signals, reports) are causally generated by its internal states in ways that go beyond simple input-output mappings.

Relation to IIT formalism: This concept relates to the intrinsicality postulate of IIT, which requires that cause-effect power be exerted by the system "within itself" (Albantakis et al., 2023). IC provides a way to discuss whether a system's outputs reflect genuine internal causal dynamics.

3.3 Self-Report Capacity (SRC)

Tentative definition: A system possesses Self-Report Capacity to the extent that it can generate signals that (a) refer to its own states, (b) vary systematically with those states, and (c) are accessible to external observers.

Relation to IIT formalism: While IIT itself does not require self-report for consciousness, the question of how Φ-structures relate to reportability is empirically important. SRC provides vocabulary for this discussion.

3.4 Persuasiveness of Self-Representation (PSR)

Tentative definition: A system exhibits Persuasiveness of Self-Representation to the extent that external observers tend to attribute internal states to it based on its outputs and behavior.

Note: PSR is explicitly observer-relative. It does not claim to measure what a system "really" has but rather what observers attribute to it—a distinction that may be useful in some research contexts.

3.5 Emphasis: These Are Only Examples

We stress again that these specific terms are illustrative. The methodological point does not depend on adopting any particular vocabulary. What matters is the general strategy: having terminology available that allows discussion of IIT's mathematical formalism without presupposing its phenomenological interpretation. Different research communities might develop different preferred terms; the specific labels are secondary.

4 Relating Operational Concepts to IIT 4.0's Framework

4.1 Preserving the Mathematical Apparatus

One advantage of supplementary operational vocabulary is that it can be used alongside IIT's existing mathematical framework without requiring any modifications to that framework. The calculations of Φ, φ_d, φ_r, and related quantities remain exactly as specified in IIT 4.0 (Albantakis et al., 2023). What changes is only how we describe what these calculations characterize—and only in contexts where the phenomenological interpretation is not the focus.

The proposed operational vocabulary allows researchers to discuss IIT 4.0's Φ-structure analysis exactly as specified, while treating the identification with phenomenal experience as a separate, interpretively open question.

4.2 Distinguishing Levels of Claim

For purposes of clarity, we might distinguish:

Level 1 (Mathematical): IIT 4.0 provides formal methods for characterizing cause-effect structures, quantifying integration, identifying maximal substrates, and unfolding Φ-structures with their constituent distinctions and relations.

Level 2 (Empirical): These methods can be applied to physical systems (neural networks, computational models, potentially biological systems as methods improve) to characterize their causal architecture.

Level 3 (Interpretive): The Φ-structure thus characterized is identical to phenomenal experience—"an experience is identical to the Φ-structure" (Albantakis et al., 2023).

Our suggestion is simply that researchers primarily working at Levels 1 and 2 might find it useful to have vocabulary that does not presuppose Level 3. This does not deny Level 3; it simply distinguishes the levels.

4.3 Neither Endorsement nor Rejection

This distinction should not be understood as either endorsing or rejecting IIT's identity claim. The claim that Φ-structure is identical to phenomenal experience is the theoretical core of IIT, and the theory stands or falls on whether this claim is ultimately vindicated. Our point is narrower: for certain purposes, discussing the formalism without constantly invoking the identity claim might be useful. Whether the identity claim is true is a separate question—one we do not address.

5 Potential Applications

5.1 Facilitating Interdisciplinary Communication

IIT's mathematical tools could potentially be useful across multiple disciplines: neuroscience, computer science, systems biology, clinical medicine. However, framing these tools as measures of "consciousness" may deter engagement from researchers wary of philosophical entanglement.

Presenting the tools as characterizing "integrated cause-effect structure" or "intrinsic causal architecture"—neutral descriptions of what the mathematics demonstrably captures—might facilitate broader uptake. Researchers could employ the tools while remaining agnostic about consciousness, examining whether the structural measures correlate with other variables of interest.

5.2 Artificial Systems

Questions about AI and consciousness have become increasingly prominent and contested. Operational vocabulary might help separate tractable questions from intractable ones:

- **Tractable**: What is the integrated cause-effect structure of this artificial system? Does it satisfy IIT's postulates? How does its Φ compare to other architectures?
- **Contested**: Is this artificial system conscious? Does it have phenomenal experience?

The first set of questions can be investigated using IIT's formal tools. The second set involves the identity claim that is IIT's interpretive core. Both sets are legitimate, but distinguishing them may prevent the contested questions from blocking progress on the tractable ones.

5.3 Clinical Applications

Clinical assessment of consciousness disorders (vegetative state, minimally conscious state, etc.) involves both scientific and ethical dimensions. Operational vocabulary might help distinguish:

- **Empirical characterization**: What does this patient's brain exhibit in terms of integrated cause-effect structure? What do Φ-related measures indicate?
- **Clinical interpretation**: What do these measures imply about the patient's experience, and what are the ethical implications?

The first question is (in principle) scientifically tractable. The second involves interpretation that goes beyond the measures themselves. Distinguishing them might promote both scientific clarity and ethical carefulness.

6 Limitations and Objections

6.1 "This Misses What IIT Is Really About"

Objection: IIT is fundamentally a theory of consciousness, not just a theory of causal structure. Discussing the formalism without the phenomenological interpretation misses the point.

Response: We agree that the identity claim is central to IIT's theoretical ambitions. Our suggestion is not that this claim should be abandoned but that, for certain purposes, it might be useful to distinguish the claim from the formalism. Researchers who accept both can continue to use the full vocabulary; researchers who wish to engage with the formalism while remaining agnostic about the identity claim would have that option.

6.2 "This Is Just Avoiding the Hard Question"

Objection: The hard problem of consciousness is hard precisely because it concerns the relationship between physical processes and subjective experience. Operational vocabulary sidesteps this question rather than addressing it.

Response: This objection has merit. Our approach does not solve the hard problem—nor does it claim to. The modest claim is that, for some research purposes, sidestepping the hard question might be methodologically appropriate, just as physicists can do productive work without resolving debates about the metaphysics of time.

6.3 "The Proposed Terms Are No Better"

Objection: Terms like "Integrated Cause-Effect Structure" are just as theory-laden as "consciousness"—they simply carry IIT's assumptions rather than folk-psychological ones.

Response: There is something to this objection. Any vocabulary carries some theoretical commitments. However, the commitments carried by operational descriptions of IIT's formalism are precisely the commitments needed to apply that formalism—nothing more. They do not carry the additional centuries of phenomenological and metaphysical baggage that "consciousness" brings.

6.4 "This Could Lead to Neglect of Subjective Experience"

Objection: Encouraging operational vocabulary might lead researchers to forget that subjective experience is what ultimately matters—and what IIT is trying to explain.

Response: This is a legitimate concern. We do not suggest that subjective experience is unimportant or that the question of its relationship to physical structure is uninteresting. We merely suggest that, for certain research purposes, distinguishing the formal characterization from the phenomenological interpretation might be useful. Whether this distinction leads to neglect or to greater clarity is an empirical question about research practice.

7 Conclusion

This paper has explored whether the conceptual vocabulary of Integrated Information Theory 4.0 (Albantakis et al., 2023) might usefully be supplemented with operationally defined terms when discussing its mathematically rigorous constructs.

We have not argued that the concept of consciousness should be abandoned, that IIT's formalism is flawed, or that the identification of Φ-structure with phenomenal experience is mistaken. Our suggestion is more modest: in strictly empirical and computational contexts, there may be advantages to distinguishing between what IIT's mathematics demonstrably characterizes—integrated cause-effect structures with their distinctions and relations—and the further claim that these structures are identical to phenomenal experience.

The specific labels we have proposed—ICES, IC, SRC, PSR—are illustrative only. Any neutral vocabulary that separates the mathematical formalism from its phenomenological interpretation would serve the same purpose. What matters is not the particular terms but the systematic separation of what IIT rigorously quantifies from the age-old interpretive leap to subjective experience.

Why does this matter? Because it seems regrettable when years of brilliant mathematical work risk being dismissed as modern scholasticism—debates about how many experiences can dance on the head of a Φ-structure—simply because the terminology used to present the work immediately activates centuries of unresolved philosophical controversy.

IIT 4.0's formalism is a genuine scientific achievement: precise, systematic, and in principle empirically applicable. This achievement deserves engagement from researchers across disciplines, including those who remain agnostic about the hard problem of consciousness. Terminological hygiene—carefully distinguishing the formalism from its phenomenological interpretation—might help ensure such engagement.

The identity claim at the heart of IIT—that Φ-structure is phenomenal experience—remains untouched by our proposal. Whether that claim is ultimately vindicated is a question for future research and philosophical reflection. In the meantime, having vocabulary that allows discussion of IIT's impressive formal machinery without presupposing the answer to that question may prove useful.

We offer these reflections in the spirit of methodological exploration, hoping they might contribute to the broader project of understanding the physical basis of mind—however that

understanding is ultimately expressed.

References

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

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

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

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

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

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

Correspondence: Boris Kriger, Institute of Integrative and Interdisciplinary Research
Email: `boriskriger@interdisciplinary-institute.org`

Acknowledgments: The author thanks colleagues for valuable discussions on methodology and terminology in consciousness research.

Declaration: The author declares no competing interests.