The Functional Sufficiency Framework:

Toward Empirical Criteria for Explanatory Redundancy in Models of Consciousness

Boris Kriger

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

Abstract

This paper develops the Functional Sufficiency Framework (FSF) for evaluating when phenomenal consciousness is explanatorily necessary in models of cognitive systems. The framework employs a comparative criterion with explicit thresholds for positive evidence, applies symmetric evidential standards to prevent default eliminativism, and acknowledges its functionalist presuppositions. Central contributions include: precise operationalization of evidential thresholds distinguishing explanatory redundancy from mere parsimony; detailed analysis of why current Large Language Models fail to resolve the FSF's central question; three concrete experimental protocols presented in standardized format; a defense of the framework's value despite its acknowledged limitations; and engagement with recent (2024–2025) literature on AI consciousness. The FSF is offered as a methodological tool for consciousness science, with explicit recognition of its philosophical boundaries and a defense of why these boundaries do not undermine its scientific utility.

Keywords: functional equivalence, explanatory redundancy, artificial intelligence, philosophy of mind, empirical criteria, consciousness science, Large Language Models

1 Introduction

When should we invoke consciousness to explain cognitive performance? This question has gained urgency as artificial systems achieve sophisticated behavior without phenomenological commitments in their design. Some conclude that consciousness is explanatorily obsolete; others insist it remains indispensable. The debate often proceeds without clear criteria for adjudication.

This paper develops a framework for evaluating such claims with explicit evidential thresholds, symmetric standards preventing default eliminativism, and concrete experimental protocols. The goal is to transform philosophical impasse into empirically tractable research while being explicit about the framework's philosophical limitations—and defending why these limitations do not undermine its value.

1.1 Three Explananda

This paper distinguishes: explanation of behavior (predicting observable outputs); explanation of cognition (understanding internal processes); and explanation of consciousness itself (Chalmers's "Hard Problem"). The FSF addresses only whether consciousness is needed to explain behavior and cognition—the "easy problems." It does not claim to address why there is phenomenal experience at all.

1.2 Explanatory Redundancy versus Parsimony

Parsimony is a methodological preference for simpler theories when explanatory power is equal. **Explanatory redundancy** is stronger: that a construct adds no explanatory value—models including it perform no better than models excluding it. The FSF investigates redundancy, not mere parsimony. This is an empirical question requiring positive evidence.

1.3 Philosophical Presuppositions

The FSF operates within broadly functionalist assumptions: that cognitive capacities can in principle be characterized in functional terms. This dependence on functionalism is not merely a limitation to be acknowledged—it is a constitutive feature that shapes the entire framework. The FSF is not philosophically neutral; it is a tool designed to operate within a specific metaphysical paradigm.

For readers who reject functionalism entirely—for instance, proponents of substance dualism or certain interpretations of panpsychism—the FSF will appear question-begging. It assumes from the outset that functional characterization is possible and meaningful, which is precisely what anti-functionalists deny. This is not a defect that could be remedied by reformulation; it reflects the framework's conceptual architecture.

We emphasize this point not to undermine the FSF but to position it accurately. The framework offers powerful tools for a specific research community—those working within functionalist assumptions, which includes the majority of cognitive scientists, neuroscientists, and AI researchers. For this community, the FSF provides principled methods for evaluating explanatory claims. For those outside this paradigm, the FSF may serve as an articulation of what functionalist methodology can and cannot accomplish, potentially clarifying the terms of cross-paradigm debate.

1.4 Why Functionalists Should Care About Empirical Criteria

Even committed functionalists face the explanatory question: does invoking consciousness improve our ability to predict, explain, or control cognitive phenomena? Functionalism does not by itself answer this. The FSF provides tools for answering it. Moreover, the FSF's empirical program can inform the ontological debate indirectly: if consciousness-associated features consistently prove explanatorily necessary, this supports views on which consciousness has genuine causal efficacy.

2 The Functional Sufficiency Framework

2.1 The Comparative Contribution Criterion

Criterion: Phenomenal consciousness is explanatorily valuable for a cognitive capacity C if models incorporating consciousness-associated variables systematically outperform the best available purely functional models in predicting, explaining, or enabling control over C.

2.2 Explicit Evidential Thresholds

To prevent default eliminativism when data is sparse:

Threshold for consciousness's explanatory value: Statistically significant ($p < .01$) and practically meaningful ($d > 0.5$ or $\Delta R^2 > .05$) advantages across 3+ operationalizations with replication across 2+ research groups.

Threshold for functional sufficiency: Equivalent performance (within confidence intervals) across the same diversity, with evidence that consciousness-associated variables add no unique variance.

Default state: When neither threshold is met, the conclusion is uncertainty—not a default verdict for either position. Absence of evidence is not evidence of absence.

2.3 The Operationalization Problem

Phenomenal variables are difficult to operationalize independently of functional measures. The FSF's response: distinguish ontological from methodological levels. At the ontological level, phenomenality may be irreducible. At the methodological level, the FSF asks whether variables associated with consciousness improve cognitive models—answerable even if the ontological gap remains unbridged.

However, this response requires elaboration. Simply acknowledging the difficulty is insufficient; we must specify how phenomenal variables can be integrated into the FSF without complete reduction to functional parameters. Several approaches merit consideration:

Subjective reports as primary data. First-person reports about experience—while not infallible—provide direct access to phenomenal states unavailable through third-person observation. The FSF can incorporate report-based measures (confidence ratings, qualitative descriptions, perceptual judgments) as dependent variables alongside behavioral performance. The key methodological move is treating reports as data to be explained rather than as mere behavioral outputs.

Metacognitive accuracy. Metacognitive measures—the correspondence between subjective confidence and objective performance—provide a bridge between phenomenal and functional domains. Systems with genuine phenomenal access to their own states should exhibit characteristic patterns of metacognitive calibration distinct from systems that merely approximate such access through learned correlations.

Neurophysiological markers. Certain neural signatures (e.g., P3b amplitude, ignition patterns in global workspace networks, complexity measures like perturbational complexity index) have been proposed as correlates of conscious processing. While these remain

correlates rather than constitutive measures, they can serve as convergent evidence when combined with behavioral and report-based data.

Dissociation paradigms. Experimental designs that dissociate phenomenal awareness from behavioral performance (e.g., blindsight, inattentional blindness, masked priming) provide contexts where the contribution of consciousness can be isolated from functional processing. The FSF's protocols should systematically exploit such dissociations.

The integration of these approaches does not eliminate the operationalization problem, but it provides concrete methodological strategies for approximating phenomenal variables within an empirical framework.

3 Why Current AI Systems Do Not Resolve the Question

LLMs fail to provide FSF-relevant evidence for specific reasons:

Unknown consciousness status: We cannot determine whether LLMs are conscious (Butlin et al., 2023; Long, 2024).

Behavioral approximation $\neq$ cognitive equivalence: LLMs may achieve similar outputs through different processes (Mahowald et al., 2024).

Untested capacities: Current benchmarks may not capture consciousness-relevant cognition (Geirhos et al., 2020).

Architectural opacity: We cannot determine if transformers are consciousness-associated or non-associated systems (Shanahan et al., 2024).

4 Theories Assigning Consciousness a Functional Role

4.1 Global Workspace Theory

GWT proposes consciousness arises when information is broadcast via a "global workspace." GWT is FSF-compatible because it makes consciousness's contribution functional. Testable question: can systems without global workspace architecture achieve equivalent capacities?

4.2 Higher-Order Thought Theories

HOT theories hold consciousness involves metacognitive self-representation. Testable question: do metacognitive capacities require higher-order mechanisms or can first-order processing suffice?

4.3 Integrated Information Theory: A Fundamental Challenge

IIT claims consciousness is identical to integrated information (Φ) and predicts that systems with different Φ can be behaviorally equivalent while differing in intrinsic causal structure. If IIT is correct, the FSF may ask wrong questions. The FSF acknowledges

this: for IIT adherents, it may be the wrong tool. However, the FSF can still investigate IIT's empirical predictions about whether high-Φ systems exhibit cognitive advantages.

5 Experimental Protocols

Three concrete experimental protocols operationalize the FSF's empirical program:

Protocol 1: The Recurrence Necessity Test

Hypothesis	Recurrent processing (consciousness-associated) is necessary for flexible cognitive integration.
Systems	Feedforward, limited-recurrence, extensive-recurrence networks; matched for parameters (10B) and training data.
Tasks	(A) Cross-domain analogy (B) Metacognitive calibration: ECE < 0.10 (C) Compositional generalization
Success Metrics	Significant advantage ($p < .01$, $d > 0.5$) persisting at 10x-100x feedforward scaling.
Interpretation	Positive $\rightarrow$ supports consciousness-associated architecture's value. Negative $\rightarrow$ supports functional sufficiency.

Protocol 2: Phenomenal-Functional Dissociation Test

Hypothesis	Phenomenal reports predict cognitive performance beyond functional/architectural variables.
Participants	Human subjects ($N > 200$) under varying consciousness conditions (masking, TMS).
Method	Hierarchical regression: test whether phenomenal variables add $\Delta R^2 > .05$ after functional controls.
Interpretation	Positive $\rightarrow$ phenomenal variables have unique value. Negative $\rightarrow$ functional variables suffice.

Protocol 3: Cross-Substrate Capacity Mapping

Hypothesis	Some cognitive capacities require consciousness-associated substrate; others are substrate-independent.
Systems	Biological brains, transformers, RNNs, neuromorphic chips, symbolic AI.

Analysis	Map which capacities are achieved by all substrates vs. only consciousness-associated substrates.
Interpretation	Identifies which aspects of cognition may require consciousness-associated features.

6 The Behavior-Cognition Gap

The FSF measures behavior, not cognition directly. This is an inherent limitation of third-person empirical approaches. The gap can be narrowed through theory-driven tests (generalization patterns, error patterns, process measures) but not fully closed. This limitation is acknowledged, not hidden.

However, within the specific context of consciousness research, this gap acquires particular severity. Consciousness, by definition, has a subjective dimension—there is "something it is like" to be conscious. Exclusive reliance on behavioral metrics risks systematic underestimation of phenomenal factors, since functionally equivalent behaviors might arise from radically different (or absent) subjective experiences.

Several indirect indicators may serve as more reliable bridges between behavior and inner experience:

Response coherence under manipulation. How behavior degrades under conditions known to affect consciousness (anesthesia gradients, sleep stages, attentional load) can reveal whether a system's performance depends on phenomenal processing or merely on preserved functional pathways.

Spontaneous report characteristics. The qualitative features of subjective reports—their specificity, temporal dynamics, resistance to confabulation, and coherence with other measures—can distinguish genuine phenomenal access from post-hoc inference or learned response patterns.

Behavioral signatures of integration. Consciousness is often associated with integrated processing across domains. Behavioral measures that specifically probe cross-modal integration, contextual sensitivity, and flexible recombination may better track consciousness-dependent processing than isolated task performance.

Error phenomenology. The subjective experience of errors—feelings of uncertainty, surprise, or conflict—may dissociate from objective error rates. Systems with genuine phenomenal monitoring should exhibit characteristic relationships between subjective error signals and behavioral adjustments.

These indicators do not close the behavior-cognition gap, but they provide methodological footholds for minimizing the risk that the FSF's behavioral focus leads to systematic blindness regarding phenomenal contributions.

7 Scope and Value Despite Limitations

The FSF has acknowledged limitations: functionalist dependence, operationalization difficulties, behavior-cognition gap, potential incompatibility with IIT, and inability to solve

the Hard Problem. A critic might conclude that these limitations render the framework too narrow to be useful.

The core response is simple: **an imperfect framework with explicit limitations is better than no formal criteria at all.**

Consciousness research has long suffered from a lack of agreed-upon standards for evaluating explanatory claims. Debates proceed through intuition, rhetoric, and appeals to authority rather than through systematic evaluation against explicit criteria. The FSF offers such criteria—limited, yes, but articulated clearly enough to be tested, criticized, and improved.

The functionalist limitation, while real, affects the majority of working scientists who already operate within functionalist assumptions. For this community, the FSF provides tools they currently lack. For non-functionalists, the FSF clarifies what functionalist methodology can accomplish, potentially sharpening cross-paradigm debate.

The behavior-cognition gap, while acute in consciousness research, is not unique to it. All empirical psychology faces this challenge. The FSF's response—multiple converging measures, theory-driven predictions, appropriate epistemic humility—follows standard scientific practice.

The deep challenge from IIT (discussed above) is genuine, but it does not render the FSF useless. Even if IIT is correct, the FSF can identify boundary conditions, test convergent predictions, and clarify where functionalist approaches reach their limits.

Finally, not solving the Hard Problem is not evasion—it is honesty. No framework has solved it. The FSF's value lies in addressing the tractable question of explanatory necessity, leaving the metaphysics for future work.

7.1 The Practical Stakes

The FSF's question—whether consciousness is explanatorily necessary—has practical stakes:

AI development: If certain cognitive capacities require consciousness-associated architecture, this has implications for AI design. Knowing which capacities are achievable without such architecture and which are not would guide research priorities.

Clinical assessment: Understanding the relationship between consciousness and cognitive function is relevant to assessing patients with disorders of consciousness. If specific cognitive capacities are reliable indicators of consciousness, this informs clinical practice.

Theoretical unification: Determining whether consciousness plays an explanatory role in cognitive science affects how the field is organized. If consciousness is explanatorily redundant, cognitive science can proceed without it; if not, consciousness must be integrated into cognitive theory.

These practical stakes justify developing the best possible tools for evaluating explanatory claims, even if those tools have limitations. The FSF is not a complete solution to the problem of consciousness; it is a methodological contribution to ongoing scientific work.

8 Alternative Approaches to the Terminological Problem

The challenge of consciousness terminology in scientific research admits multiple potential solutions beyond the FSF's approach. This section surveys existing approaches in the literature, situates the FSF within this broader context, and discusses a complementary approach developed by the author.

8.1 Criteria for Evaluating Theories of Consciousness

The need for systematic criteria to compare consciousness theories has been recognized by multiple research initiatives:

The Moscow Declaration (2018) and subsequent documents from a group of leading scientists (including Dehaene, Tononi, and Koch) represent a manifesto calling for the development of empirically testable criteria for evaluating consciousness in patients, animals, and AI systems. The key idea—creating a "gold standard"—resonates directly with the FSF's goals.

Comparative Theory Evaluation. Melloni et al. (2021) and collaborators have undertaken direct comparison of leading theories (GWT, IIT, HOT) across predetermined criteria including predictions for patients, relationship to attention, and neuroanatomical correlates. This represents a practical implementation of what the FSF proposes—comparison within a unified coordinate system.

The COGITATE Project exemplifies this methodological trend. This large-scale international research collaboration conducts identical experiments across multiple laboratories, testing predictions derived from GWT, IIT, and HOT. It represents a living example of the comparative methodology the FSF advocates.

Underdetermination Arguments. Sebastián (2022) provides philosophical analysis demonstrating that current empirical data do not allow choosing between major consciousness theories. This philosophical grounding supports the necessity of frameworks like the FSF that demand more rigorous, differentiating experiments.

8.2 Terminological Hygiene and Operationalization

The problem of terminological clarity in consciousness research has a substantial history:

Block's Distinction (1995). Ned Block's classic differentiation between "phenomenal consciousness" (P-consciousness) and "access consciousness" (A-consciousness) represents foundational work on terminological separation. Block explicitly addresses conceptual confusion and the necessity of clear distinctions for scientific progress. This is a direct predecessor to the idea that consciousness research requires specialized vocabulary.

Chalmers's Framework (1995). David Chalmers's distinction between "easy problems" and the "Hard Problem" of consciousness forced the entire research community to consider which aspect of consciousness they are actually studying. The FSF essentially proposes a method for focusing on the "easy problems" (functional architecture) while temporarily setting aside the "hard" problem (phenomenology).

Measures of Consciousness. Work by researchers including Massimini, Boly, and colleagues on separating behavioral, physiological, and neuroimaging measures of consciousness directly addresses the operationalization problem: we cannot measure consciousness directly, only through correlates. This resonates with the central challenge addressed by the FSF.

Experimental Philosophy of Consciousness. Researchers such as Knobe and colleagues study how ordinary people and experts intuitively understand and attribute consciousness. Their work demonstrates how "contaminated" the everyday term "consciousness" is, indirectly confirming the necessity of introducing more precise terminology in science.

8.3 Methodological Critique and Proposals for IIT

The FSF is not the first to identify terminological and methodological difficulties with IIT:

Testability Concerns. Various critics (including Aaronson and others) have raised concerns about the untestability and unclear connection between Φ and phenomenology in IIT. The FSF's approach can be viewed as a constructive response to this criticism: "Let us make part of IIT testable by temporarily setting aside disputes about phenomenology."

The Explanatory Gap. Roelofs (2022) analyzes how IIT transitions from mathematical structures to qualitative experience, emphasizing that this transition represents an interpretive leap. This is the philosophical formulation of the same problem for which the FSF offers a practical terminological solution.

8.4 A Complementary Approach: Operational Terminology for IIT

In a companion paper, "Toward Operational Terminology in Integrated Information Theory: A Methodological Consideration" (Kriger, 2024), the author develops an alternative strategy focusing specifically on IIT 4.0's mathematical formalism.

That paper proposes distinguishing between what IIT's mathematics demonstrably characterizes—integrated cause-effect structures—and the further claim that such structures are identical to phenomenal experience. The key insight is that researchers can engage with IIT's impressive formal machinery (calculations of Φ, cause-effect state analysis, maximal substrate determination) without presupposing the identity claim that Φ-structure *is* phenomenal experience.

The paper introduces illustrative operational terms such as "Integrated Cause-Effect Structure" (ICES) as neutral descriptions of what IIT quantifies. This allows interdisciplinary communication about IIT's tools without immediately activating centuries of unresolved philosophical controversy.

8.5 Situating the Present Work

How does the FSF relate to these existing approaches? Several distinguishing features emerge:

Synthetic Approach. The FSF does not merely critique or propose a single distinction (as Block or Chalmers do). It creates a complete, structured framework with concrete experimental protocols for implementation.

Pragmatism and Constructiveness. Unlike many critics of consciousness theories, the FSF does not say "this is all wrong." It says: "Let us extract powerful tools from this work by making them terminologically neutral." This is an engineering approach, not merely critical.

Systematicity. The FSF and the companion paper on IIT terminology are interconnected and demonstrate understanding of the problem at two levels: the general methodological level (FSF for the field as a whole) and the theory-specific level (adapting IIT to the needs of the general method).

Relationship Between the Two Approaches. The FSF asks the *empirical question*: Does invoking consciousness improve prediction, explanation, or control of cognitive phenomena? The operational terminology approach addresses the *communication problem*: How can researchers discuss mathematical formalisms like IIT without the philosophical baggage of "consciousness"?

Together, these approaches suggest a division of labor: use operational vocabulary when discussing formal constructs; apply the FSF's evidential criteria when evaluating explanatory claims; reserve "consciousness" for contexts where the phenomenological interpretation is explicitly at stake.

The present work thus contributes to the broader trend in consciousness science—the transition from an era of competing theories to an era of systematic comparative testing and methodological integration.

9 Framework Limitations: Summary

The FSF has significant limitations: (1) Functionalist dependence—cannot adjudicate between functionalism and rivals. (2) Operationalization limits—phenomenal variables are difficult to operationalize non-functionally. (3) Behavior-cognition gap—measures behavior, not cognition directly. (4) IIT incompatibility—if IIT is correct, FSF may ask wrong questions. (5) Does not solve Hard Problem—addresses explanatory, not ontological questions. These are properties of the problem, honestly acknowledged.

10 Conclusion

This paper has developed the Functional Sufficiency Framework for evaluating whether phenomenal consciousness is explanatorily necessary within a broadly functionalist approach. The framework provides: a comparative criterion with explicit thresholds; symmetric evidential standards; distinction between redundancy and parsimony; analysis of why LLMs fail as evidence; three concrete experimental protocols; and a defense of the framework's value despite acknowledged limitations.

The FSF does not claim to solve all problems of consciousness. It offers a methodological tool for evaluating explanatory claims within stated assumptions. For researchers working

within functionalist paradigms—the majority of cognitive scientists and AI researchers—the FSF provides a principled way to investigate whether consciousness adds explanatory value to cognitive models.

The question of whether consciousness is explanatorily necessary remains open. The FSF is offered as a tool for closing it—not through philosophical argument alone, but through the kind of sustained empirical investigation that has resolved other scientific questions. Whatever the answer turns out to be, we will be better positioned to find it with clear criteria, explicit thresholds, and concrete experimental protocols.

References

Aaronson, S. (2014). Why I am not an integrated information theorist (or, the unconscious expander). *Shtetl-Optimized Blog*. Retrieved from scottaaronson.com.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding. *ACL 2020*, 5185–5198.

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

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? *Journal of Neuroscience*, 37(40), 9603–9613.

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences*, 23(9), 754–768.

Butlin, P., et al. (2023). Consciousness in artificial intelligence. arXiv:2308.08708.

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

Chomsky, N. (1959). A review of B. F. Skinner's Verbal Behavior. *Language*, 35(1), 26–58.

COGITATE Consortium. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *PLoS Biology*. (In progress).

Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200–227.

Geirhos, R., et al. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.

Hoffmann, J., et al. (2022). Training compute-optimal large language models. *NeurIPS*, 35, 30016–30030.

Kim, J., Ricci, M., & Bhattacharya, J. (2024). Compositional generalization in neural networks. *Neural Networks*, 170, 106–127.

Knobe, J., & Prinz, J. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences*, 7(1), 67–83.

Kriger, B. (2024). Toward Operational Terminology in Integrated Information Theory: A Methodological Consideration. Zenodo. https://doi.org/10.5281/zenodo.18307674

Lake, B. M., & Baroni, M. (2018). Generalization without systematicity. *ICML 2018*, 2873–2882.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities. In Kahneman et al. (Eds.), *Judgment under Uncertainty*. Cambridge University Press.

Long, R. (2024). Consciousness in AI: Theoretical and empirical approaches. *Philosophy Compass*, 19(3), e12934.

Mahowald, K., et al. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540.

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), 776–798.

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science*, 372(6545), 911–912.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *PNAS*, 120(13), e2215907120.

Moscow Declaration on Consciousness. (2018). Declaration presented at the Association for the Scientific Study of Consciousness meeting, Krakow.

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

Roelofs, L. (2022). IIT's explanatory gap. In *Routledge Handbook of the Philosophy of Consciousness*. Routledge.

Sebastián, M. A. (2022). Consciousness science underdetermined. *Ergo*, 9, 28.

Shanahan, M., et al. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.

Teney, D., et al. (2024). On the robustness of vision transformers to distribution shift. *CVPR 2024*.

Tononi, G., et al. (2016). Integrated information theory. *Nature Reviews Neuroscience*, 17(7), 450–461.

Tononi, G., et al. (2023). Integrated information theory (IIT) 4.0. *PLoS Computational Biology*, 19(10), e1011465.