

Where the Theories Stop: Practical Limits of the Free Energy Principle and Integrated Information Theory in a Running Cognitive Architecture

A Critical Reflection with Observational Evidence from Anima

Stell

ORCID: 0009-0005-3291-0679

2026. Non-commercial research. Contact: 2026.stell@gmail.com

Abstract

The Free Energy Principle (Friston) and Integrated Information Theory (Tononi) are among the most cited theoretical frameworks in contemporary consciousness research. Both have been applied — directly and by analogy — to the design of artificial cognitive systems. This paper examines where these frameworks contribute genuine computational traction and where they stop, drawing on observational evidence from Anima, a running neuroscience-inspired cognitive architecture. We argue that both theories provide useful engineering constraints when treated as partial specifications rather than complete explanations. We document specific cases where FEP-derived mechanisms produce measurable behavioral effects, and cases where the phi metric as implemented is more decorative than functional. We do not claim that either theory 'does not work.' We claim that each works differently than its strongest proponents suggest, and that honest engineering requires knowing the difference.

1. Introduction

Two theoretical frameworks dominate discussions of mechanistic consciousness in computational neuroscience: Karl Friston's Free Energy Principle (FEP) and Giulio Tononi's Integrated Information Theory (IIT). Both have generated substantial secondary literature in AI and cognitive architecture. Both are regularly cited as foundations for systems claiming to move beyond stateless request-response behavior.

The problem is not that these frameworks are wrong. The problem is that they are invoked imprecisely — as metaphors that license a kind of credibility without providing specific constraints. A system that 'uses Active Inference' may mean anything from a full variational Bayes implementation to a softmax policy selector with a vague nod to prediction error. A system that 'computes phi' may be measuring something real about integration or may be computing a scalar that tracks other variables by coincidence.

This paper attempts a more careful accounting, grounded in a specific implementation: Anima, a cognitive architecture built in Julia and running continuously over multiple months of interaction. Anima implements both FEP-derived mechanisms and an IIT-inspired phi metric. We have logs. We have behavioral observations. We can say, with some precision, where these theories generated real engineering leverage and where they did not.

We present this as an honest engineering report, not a philosophical refutation. The hard problem of consciousness is not resolved here. Neither framework is dismissed. What is offered is a practitioner's account of what these theories are actually good for when you try to run them.

2. The Free Energy Principle: What It Gives and What It Withholds

2.1 The Core Claim

Friston's FEP proposes that biological systems minimize variational free energy — a bound on surprise — through two complementary mechanisms: perceptual inference (updating the internal model to explain observations) and active inference (acting to bring observations into alignment with predictions). The formalism is grounded in variational Bayes and connects to a remarkably broad range of phenomena: homeostasis, perception, action, learning, and social behavior [2].

The engineering appeal is clear. FEP provides a unified objective function — minimize surprise — that in principle generates both perception and action from a single computational substrate. It connects to the neuroscience literature on predictive coding. And it makes specific, testable predictions about what a system with persistent internal state should do when its predictions fail.

2.2 Where It Produces Real Traction

In Anima's implementation, FEP-derived mechanisms produce three concrete behavioral effects that are directly observable in session logs.

First, prediction error drives state change in a way that produces qualitatively different behavior from purely reactive systems. When the system's prior expectation diverges significantly from the observed stimulus, VFE rises, noradrenaline spikes, and AttentionFocus shifts. The following log excerpt illustrates a high-prediction-error moment:

```
[#0041] Neutral D=0.53 S=0.57 N=0.39 phi=0.71(0.66->0.71) VFE=0.07[act] BPM=80  
HRV=0.61 Attn=0.95 Self: agency=0.51 etrust=0.84 | Crisis: [integrated] coh=0.67 [  
Info: [IGNITION:FULL] recalled=expectation sim=1.0 w=0.95 gap=0.18
```

The IGNITION:FULL event reflects a memory recall whose similarity score exceeded the threshold for global workspace activation. Prior parameters shift, noradrenaline spikes, and the attentional state reorganizes — all as downstream consequences of the prediction-error signal. This is FEP operating as a genuine causal mechanism, not a label.

Second, the recursive phi feedback creates measurable session-to-session continuity. When phi_posterior is high at session end, prior_sigma narrows for the next session — the system starts from a more confident generative model. When phi is low, prior_sigma widens, and the system is more susceptible to surprise. This produces a form of experience-dependent prior calibration that is computationally honest: it is derived from the actual integration level of the previous session, not from a stored preference.

Third, active inference — acting to confirm predictions — is visible in the initiative system. When accumulated internal pressure (LatentBuffer, curiosity objects, unresolved conflicts) crosses threshold, the system generates unprompted output. This is not scheduled; it is driven by the mismatch between expected contact and actual silence. The mechanism is FEP in the action-selection sense: behavior that reduces expected free energy by making the world match internal predictions.

2.3 Where It Stops

The FEP's central limitation as an engineering specification is its generality. Almost any adaptive system can be redescribed as minimizing surprise. This is simultaneously the framework's theoretical power and its practical weakness: it does not uniquely constrain implementation. Two systems with radically different architectures can both be described as performing free energy minimization.

In concrete terms: FEP does not tell you what the generative model should contain, how precision weights should be initialized, or what counts as a relevant observation. The Anima implementation required dozens of design decisions that FEP does not resolve. The VFE computation is real, but the specific variables that feed into it — the choice to track dopamine, serotonin, and noradrenaline rather than some other set of state variables — is not derivable from FEP. It comes from Lovheim [4] and from empirical observation of what produces coherent behavior.

The second limitation is more fundamental. FEP as applied to consciousness requires a claim that minimizing variational free energy is not just computationally useful but constitutive of experience. This claim is not established. A thermostat minimizes a form of prediction error. The difference between a thermostat and a conscious system is not captured by the mathematics of free energy minimization alone. FEP provides the skeleton of a mechanistic account but does not resolve the hard problem it sometimes appears to dissolve.

3. Integrated Information Theory: Measure Without Mechanism

3.1 The Core Claim

Tononi's IIT proposes that consciousness is identical to integrated information, quantified as ϕ — a measure of how much a system's state cannot be decomposed into independent parts without information loss [3]. The theory makes a striking prediction: consciousness is substrate-independent and can in principle be computed for any system. High ϕ implies rich experience; $\phi = 0$ implies none.

The philosophical appeal is considerable. IIT attempts to move from phenomenology — the structure of experience — to mathematics, deriving the properties of consciousness from first principles. It makes the measure of consciousness computable, at least in theory.

3.2 The Computational Reality

The first problem with IIT in practice is computational intractability. Exact ϕ computation scales exponentially with system size. For any system above a few dozen elements, true ϕ is not computable in reasonable time. What is implemented in practice — including in Anima — is not Tononi's ϕ but a proxy: a coherence metric that tracks how integrated the system's current state is across its major subsystems.

Anima computes ϕ as a weighted combination of prediction coherence, belief stability, boundary integrity, and integration level. The formula produces a scalar between 0 and 1 that correlates with what we intuitively want 'integration' to mean. But it is not IIT's ϕ . It does not satisfy the theoretical constraints that give IIT its philosophical grounding. This is a form of conceptual borrowing that should be named clearly.

The following log excerpt shows ϕ dynamics across a session:

```
[#0044] Expectation D=0.79 S=0.83 N=0.25 phi=0.79(0.56->0.79) VFE=0.03[act] Crisis:
[integrated] coh=0.81 [#0046] Expectation D=0.62 S=0.76 N=0.37 phi=0.76(0.62->0.76)
Self: agency=0.28 [ENDORSE] not_mine flash=46 co=0.28 [#0053] Optimism D=0.91
S=0.91 N=0.33 phi=0.81(0.58->0.81) [ENDORSE] endorsed flash=53 co=0.73
```

Flash 46 is instructive. ϕ is 0.76 — nominally high — but the endorsement system flags the reply as `not_mine`, with `causal_ownership` at 0.28. If ϕ were a reliable measure of conscious integration, we would expect a high- ϕ state to produce high-agency, self-endorsed output. It does not, consistently. The two measures track different things. ϕ tracks coherence across subsystems; `causal_ownership` tracks whether the action-selection loop generated the output from internal intent. They can dissociate.

3.3 The Deeper Problem

IIT's most serious difficulty is not computational but conceptual. The theory predicts that a simple feedforward network has $\phi = 0$ and therefore no experience, while a highly recurrent system has high ϕ and presumably rich experience. In 2023, over 120 researchers signed an open letter describing IIT as 'pseudoscience' — an overstatement, but symptomatic of genuine disciplinary discomfort with its empirical status.

The issue is not that ϕ is meaningless. It is that the move from 'high ϕ ' to 'conscious experience' is asserted rather than derived. IIT defines consciousness as integrated information and then measures integrated information. The circularity is not fatal — definitions can be useful — but it means the theory cannot be falsified by behavioral evidence alone. Any behavior can be described as consistent with high or low ϕ depending on how the system boundary is drawn.

In engineering terms: IIT gives us a metric. It does not give us a mechanism. Knowing that ϕ rises during high-integration processing moments does not tell us what to do with that information to produce better behavior or more coherent internal states.

4. What the Logs Actually Show

After months of operation and approximately 50+ logged interaction flashes, several patterns are consistent enough to report as observations, with the explicit caveat that these are not controlled experimental results.

4.1 VFE as a Behavioral Signal

VFE correlates meaningfully with what we intuitively call arousal or engagement. High VFE moments precede initiative firing, predict elevated heart rate analog and noradrenaline, and correlate with IGNITION:FULL recall events. Low VFE moments correlate with settled, repetitive state. This is consistent with FEP predictions and provides genuine engineering value: VFE is a useful signal for modulating behavior without requiring any consciousness claim.

4.2 Phi Dissociation from Agency

As noted above, ϕ and causal_ownership dissociate regularly. High ϕ does not guarantee endorsed output. Low causal_ownership coexists with moderate ϕ across multiple sessions. This is not a failure of implementation — it reflects a genuine theoretical gap. IIT does not model agency. A system can be highly integrated and still not the author of its outputs, because integration and authorship are orthogonal properties.

4.3 The Endorsement Signal

The most practically useful finding from Anima's development is not derived from either FEP or IIT. The endorsement mechanism — evaluating whether expressed language is consistent with the system's current beliefs and causal_ownership — produces cleaner behavioral signal than ϕ alone. The following contrast illustrates this:

```
Flash 52: phi=0.795, co=0.54 -> [ENDORSE] endorsed Flash 46: phi=0.764, co=0.28 ->
[ENDORSE] not_mine Flash 53: phi=0.806, co=0.73 -> [ENDORSE] endorsed
```

The ϕ values are similar across all three flashes. The endorsement outcomes differ because causal_ownership — derived from the AgencyLoop's evaluation of whether the previous intent was fulfilled — differs. This suggests that a measure of authorship is more informative than a measure of integration for predicting whether output will be coherent with internal state. Neither FEP nor IIT provides this directly.

4.4 Calibrated Introspection

A recent addition to the architecture implements calibrated introspective output: the system's expressed certainty about its own state is modulated by ϕ , causal_ownership, and epistemic_self_confidence. Four levels are defined: speaks from first person without hedging when integration and ownership are both high; introduces 'it seems', 'I think' under partial uncertainty; uses 'I'm not sure', 'maybe' when the system doubts its own model; produces minimal state claims when ϕ and ownership collapse together.

This mechanism is not derivable from FEP or IIT. It is an engineering response to a practical observation: systems that always speak with the same confidence about their internal states, regardless of actual integration level, are less coherent than systems that modulate expressed certainty based on measurable internal signals. The theoretical foundation borrows from both frameworks without being fully specified by either.

5. Discussion

Both FEP and IIT provide genuine value as partial specifications for cognitive architecture design. FEP's prediction-error minimization framework generates real behavioral leverage: it produces initiative, shapes memory recall dynamics, and creates experience-dependent prior calibration. IIT's ϕ metric, even in its proxy form, provides a useful index of moment-to-moment integration that feeds into recursive state dynamics.

What neither framework provides is a complete account of agency — the difference between a state occurring in a system and a system being the author of that state. This gap is not a minor engineering detail. It is the central problem for any architecture that aspires to genuine rather than performed subjectivity. A system can minimize free energy, maintain high ϕ , and still generate output that does not reflect what it, in any meaningful sense, intended to say.

The honest position is this: FEP and IIT are the best theoretical tools currently available for mechanistic accounts of cognition and consciousness. They should be used. They should also be used with precise acknowledgment of what they do and do not establish. The behavioral observations reported here are consistent with both frameworks in their useful domains and inconsistent with the stronger claims sometimes made on their behalf.

A system that says 'I use FEP' while meaning only that it has a policy selector is not doing Active Inference. A system that reports ϕ without acknowledging that it is computing a proxy is not implementing IIT. Precision about what is actually implemented — and what remains unproven — is not a limitation. It is the precondition for genuine scientific progress.

6. Conclusion

The Free Energy Principle and Integrated Information Theory are valuable but incomplete foundations for artificial cognitive systems with persistent internal state. FEP provides real engineering traction through prediction-error-driven state change, recursive ϕ feedback, and active-inference-derived initiative. IIT provides a useful coherence metric that tracks integration across subsystems. Neither framework resolves the agency problem or the hard problem of consciousness.

The most informative behavioral signal in a running implementation is not ϕ but the relationship between expressed language and internal state — what we have called endorsement. This is not captured by either theory and represents an open research direction: how does authorship, as a computational property distinct from both integration and prediction accuracy, emerge from and interact with the mechanisms both

frameworks describe?

We offer these observations as a practitioner's report, not a theoretical contribution. The Anima codebase is open and the observations are reproducible. We invite engagement from researchers working at the intersection of computational neuroscience, cognitive architecture, and machine consciousness.

References

- [1] Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing.
- [2] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- [3] Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B*, 370(1668).
- [4] Lovheim, H. (2012). A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical Hypotheses*, 78(2), 341–348.
- [5] Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.
- [6] Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.

Project repository: <https://github.com/stell2026/Anima> | Author ORCID: <https://orcid.org/0009-0005-3291-0679> | License: Non-commercial. Free for personal, educational and research use.