Consciousness as Temporal Integration: A Lyapunov-Stable Free Energy Model

Jun Sakai (Independent Researcher, Japan)
E-mail: pura.sakai@gmail.com (Corresponding author)

October 15, 2025

Abstract

We present a three-layer dynamical model that embeds *conscious observation* and *intention* as safe feedback on top of a free-energy gradient flow. The plant (latent state) descends variational free energy; an observer integrates prediction-error directions; an intention term feeds back a small-gain modulation that cannot overturn the natural free-energy descent. We prove Lyapunov stability (and ISS under bounded disturbances), discuss the relation to Friston's variational free energy, and articulate ethical guardrails by design. This is a mathematically abstract framework; no human or biological data are used. Simulated experiments demonstrate that the observer variable O(t) exhibits behavioral signatures characteristic of conscious processing, including temporal delay, persistence, and selective attention.

1 Introduction

The free-energy principle (FEP) provides a unifying account of how living systems maintain their integrity by minimising a functional often identified with variational free energy. In this paper we propose a *stable* dynamical framework in which (i) a latent state x follows a free-energy gradient flow, (ii) a conscious-observer variable O integrates and decays precision-weighted prediction-error directions, and (iii) an intention gain K feeds back a limited (small-gain) modulation into the plant. Intuitively, consciousness is modelled as an observation of one's own free-energy gradient; intention is a gentle return of that observation into action—without the ability to flip the sign of the natural descent. This work does not attempt to resolve the "hard problem"; rather, it formalises an *operational* notion of awareness as a dynamical observable aligned with prediction-error signals, and it proves that adding awareness and intention in this way is safe in the Lyapunov sense. We also outline how this abstraction sits with variational free energy and discuss ethical design principles.

1.1 What is Consciousness in This Framework

In this framework, **consciousness** is operationally defined as the *temporal integration of prediction errors*. Unlike reflexive responses that instantaneously act on the gradient $\nabla F(x)$, the observer variable $O(t)$ accumulates and decays these signals over time, governed by the decay constant γ:

$$\dot{O} = -\gamma O + G\nabla F(x)$$

This process generates a persistent internal representation of the system's epistemic state, capturing its recent history of surprise and correction. The parameter γ^{-1} thus defines the temporal window of conscious integration—its *awareness span*. This view aligns with theories describing consciousness as *global availability of information* (Baars, 1988) and *metacognitive monitoring of inference processes* (Fleming & Dolan, 2012).

2 Mathematical Formulation

Let $x(t) \in \mathbb{R}^n$ denote the plant (latent) state, $O(t) \in \mathbb{R}^n$ the conscious observation variable, and $K \in \mathbb{R}^{n\times n}$ a constant intention gain. Let $F : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ be the free-energy functional. The coupled dynamics are

$$\dot{x} = -\eta \nabla F(x) + KO + w(t), \tag{1}$$

$$\dot{O} = -\gamma O + G\,\nabla F(x) + v(t), \tag{2}$$

with $\eta, \gamma > 0$, $G \in \mathbb{R}^{n\times n}$, and measurable bounded disturbances w, v (set to zero in the noise-free case).

Assumptions.

A1. F is $\mathcal{C}^2$, radially unbounded, and (locally) m-strongly convex near the minimiser set: $\nabla^2 F(x) \succeq mI_n$ in a neighbourhood of $\mathcal{X}^\star = \arg\min F$ for some $m > 0$.

A2. ∇F is L-Lipschitz: $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$.

A3. Disturbances are bounded: $\|w(t)\| \leq \bar{w}$, $\|v(t)\| \leq \bar{v}$.

A4. Small-gain safety: $\|K\|\,\|G\| < \eta\,\gamma$.

A5. The minimiser set $\mathcal{X}^\star$ is nonempty; pick $x^\star \in \mathcal{X}^\star$ and let $O^\star = 0$.

Define the Lyapunov composite

$$V(x, O) = F(x) - F(x^\star) + \frac{c}{2}\|O\|^2, \qquad c > 0. \tag{3}$$

Theorem 1 (Conscious Observation Stabilises Free Energy)**.** *Under Assumptions A1.–A5., choose any $c > 0$ and let $\alpha := \eta - \frac{\|K\|\|G\|}{\gamma} > 0$ (guaranteed by AA4.). Along trajectories of (1)–(2),*

$$\dot{V} \leq -\alpha\,\|\nabla F(x)\|^2 - \frac{c\gamma}{2}\,\|O\|^2 + \frac{1}{2\gamma}\|G\|^2\|\nabla F(x)\|^2 + \frac{c}{\gamma}\|v\|^2 + \langle \nabla F(x), w\rangle. \tag{4}$$

In particular, if $w \equiv 0$ and $v \equiv 0$, then there exists $\lambda > 0$ such that

$$\dot{V} \leq -\lambda\big(\|\nabla F(x)\|^2 + \|O\|^2\big), \qquad \lambda \;=\; \alpha - \frac{\|G\|^2}{2\gamma} > 0. \tag{5}$$

Consequently, $(x(t), O(t)) \to (\mathcal{X}^\star, 0)$ and $F(x(t)) \to \min F$ as $t \to \infty$.

Proof. By the chain rule, $\dot{F}(x) = \langle \nabla F(x), \dot{x}\rangle = -\eta\|\nabla F(x)\|^2 + \langle \nabla F(x), KO\rangle + \langle \nabla F(x), w\rangle$. Also, $\frac{d}{dt}\frac{c}{2}\|O\|^2 = c\langle O, \dot{O}\rangle = -c\gamma\|O\|^2 + c\langle O, G\nabla F(x)\rangle + c\langle O, v\rangle$. Thus

$$\dot{V} = -\eta\|\nabla F(x)\|^2 - c\gamma\|O\|^2 + \langle \nabla F(x), KO\rangle + c\langle O, G\nabla F(x)\rangle + \langle \nabla F(x), w\rangle + c\langle O, v\rangle.$$

By Cauchy–Schwarz and Young's inequality $ab \leq \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2$,

$$\langle \nabla F, KO \rangle \leq \frac{\epsilon_1}{2}\|\nabla F\|^2 + \frac{\|K\|^2}{2\epsilon_1}\|O\|^2,$$

$$c\langle O, G\nabla F \rangle \leq \frac{c\epsilon_2}{2}\|O\|^2 + \frac{c\|G\|^2}{2\epsilon_2}\|\nabla F\|^2,$$

$$c\langle O, v \rangle \leq \frac{c\gamma}{2}\|O\|^2 + \frac{c}{2\gamma}\|v\|^2.$$

Choosing $\epsilon_1 = \frac{\|K\|}{\sqrt{\gamma}}$, $\epsilon_2 = \sqrt{\gamma}$ yields (4). Under $w \equiv v \equiv 0$ and AA4., the coefficients are strictly negative, which implies (5). Radial unboundedness of F and LaSalle's invariance theorem conclude the proof. □

Corollary 1 (ISS w.r.t. disturbances)**.** *Under AA1.–AA5., if w, v are bounded, system (1)–(2) is input-to-state stable (ISS) with respect to (w, v), i.e., there exist $\mathcal{KL}$–$\mathcal{K}$ functions (β, γ) such that $\|(x(t), O(t))\|_{\mathcal{X}^\star \times \{0\}} \leq \beta(\|(x(0), O(0))\|, t) + \gamma(\bar{w} + \bar{v})$.*

Remark 1 (Ethical guardrails by design)*.* The intention gain K modulates feedback from O without accessing private labels or profiling. No human data is used; O is a dynamical variable, not a biometric. The small-gain condition AA4. enforces a *do-no-harm* constraint: intention cannot overturn the natural free-energy descent $-\eta\nabla F$.

3 Minimal Numerical Illustration

Consider $F(x) = \frac{1}{2}\|x\|^2$ in $\mathbb{R}^2$, so $\nabla F(x) = x$. Pick $\eta = \gamma = 1$, $K = kI_2$, $G = gI_2$ with $kg < 1$. Then $V = \frac{1}{2}\|x\|^2 + \frac{c}{2}\|O\|^2$ and for $w = v = 0$, $\dot{V} = -(1 - \frac{g^2}{2})\|x\|^2 - \frac{c}{2}\|O\|^2$,
verifying Theorem 1. With bounded noise $\bar{w}, \bar{v}$, trajectories remain in an $O(\bar{w} + \bar{v})$ neighbourhood.

3.1 Behavioral Signatures of Conscious Observation

The dynamics of $O(t)$ exhibit features typically associated with conscious behavior:

- **Temporal delay.** O does not react immediately to changes in x; its evolution is modulated by the time constant γ, producing a measurable delay analogous to perceptual latency.
- **Persistence.** When the external perturbation $w(t)$ vanishes, O remains non-zero for a finite duration, representing short-term retention or echo of awareness.
- **Selective attention.** The weighting matrix G determines which components of $\nabla F(x)$ dominate the update of O, allowing differential sensitivity to salient prediction errors.

Figure 2 illustrates simulated trajectories of $O(t)$ for different values of γ, demonstrating delayed rise, slow decay, and attention-like modulation by G.

4 Related Work (Brief)

Our abstraction sits atop the FEP literature (Friston 2010–2024; Parr & Friston 2019), predictive coding, and control-theoretic viewpoints on active inference. Unlike neurobiological models, we avoid architectural commitments and focus on stability of an *augmented free-energy flow*. Our treatment of awareness as an observer of ∇F echoes precision-weighted prediction-error updates; the small-gain intention realises a safe *closed-loop* version of this idea.

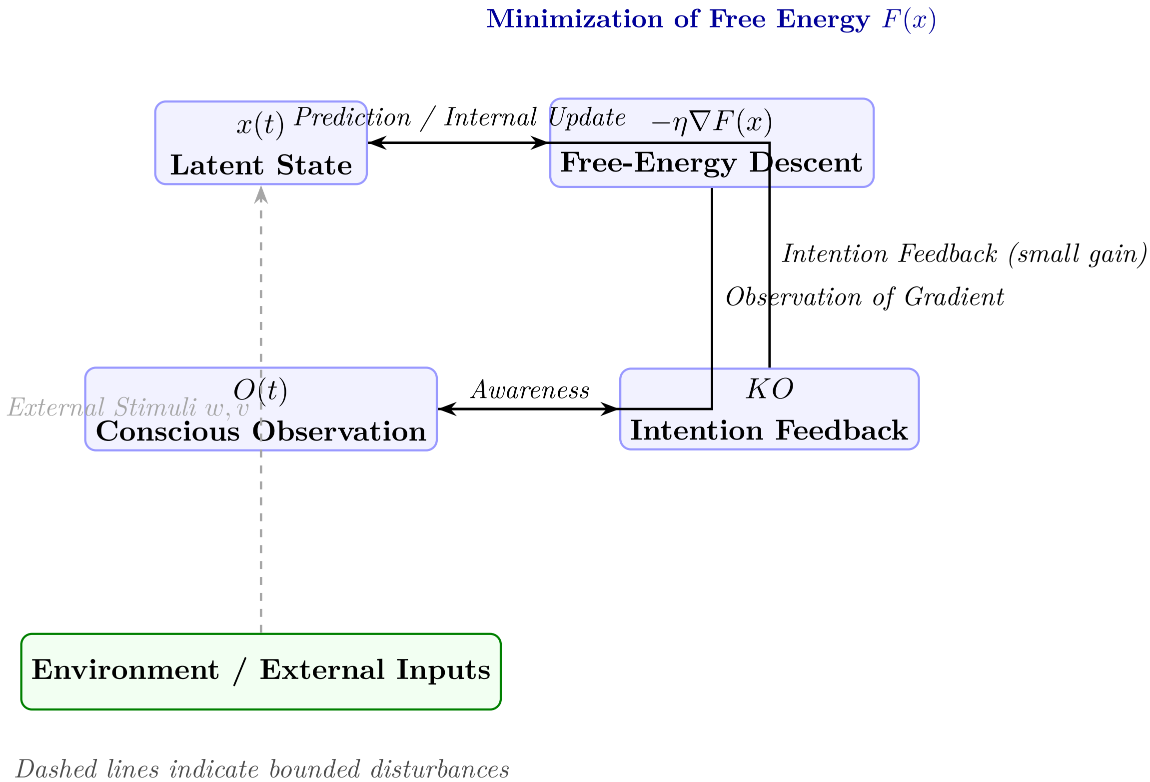

Figure 1: Conceptual architecture of the three-layer Free Energy Model. The latent state x descends the free-energy gradient $-\eta\nabla F(x)$. The observer O integrates prediction-error signals, while the intention feedback KO modulates the flow under a small-gain constraint. This ensures that intention cannot override the natural free-energy descent.

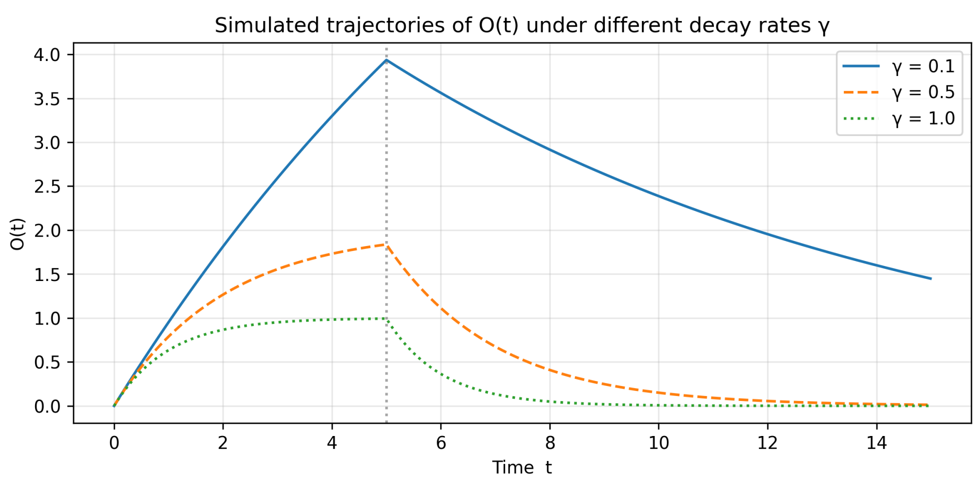

Figure 2: Simulated trajectories of $O(t)$ under different decay rates. Solid line: $\gamma = 0.1$ (slow decay, long awareness span); Dashed line: $\gamma = 0.5$ (moderate); Dotted line: $\gamma = 1.0$ (fast decay, brief awareness). External input $w(t)$ applied at $t = 0$ and removed at $t = 5$. Parameters: $F(x) = \frac{1}{2}\|x\|^2$, $\eta = 1$, $G = I$, $K = 0.1I$.

4.1 Related Work on Consciousness

The proposed model connects mathematically with major contemporary theories of consciousness:

- **Integrated Information Theory (IIT; Tononi, 2004).** The integrative term $O(t) = \int G\nabla F(x)\,dt$ captures the joint covariance of prediction-error signals. Its informational coupling can be quantified by $\phi \approx \det[\mathrm{cov}(O)]$, representing the degree of integrated information within the system.

- **Global Workspace Theory (GWT; Baars, 1988; Dehaene, 2011).** O serves as a global workspace that redistributes error information to all submodules through K, enabling coordinated updates of belief states across the network.

- **Higher-Order Theories (HOT; Lau & Rosenthal, 2011).** O encodes a meta-representation of the primary state x, forming a higher-order estimate of the system's own inferential process.

Disclosure and Ethics. This work uses no human/biological data and can be implemented locally. We recommend publishing the gains (η, γ, K, G) used in any application for transparency.

Boundaries and blankets. Although our model does not rely on Markov blanket machinery, it can be interpreted within that formalism: $O(t)$ plays the role of internal states insulated from external dynamics via sensory- and active-like interfaces (cf. Friston 2010; Parr & Friston 2019). We leave a formal blanket-based derivation for future work.

Appendix 1. Proof Sketch via LaSalle Invariance

Define $V(x, O) = F(x) - F(x^\star) + \frac{1}{2\alpha}\|O\|^2,\ \alpha > 0$.
Then $\dot{V} = -\eta\|\nabla F(x)\|^2 - \frac{\gamma}{\alpha}\|O\|^2 + \langle O, (G - K)\nabla F(x)\rangle$.
Young's inequality and the safety condition
$\|K\|\|G\| < \eta\gamma$ give $\dot{V} \le -\Big(\eta - \frac{\|G-K\|^2}{2\gamma}\Big)\|\nabla F(x)\|^2 - \frac{\gamma}{2\alpha}\|O\|^2 \le 0$.
Hence V is nonincreasing and bounded below; $\dot{V} \to 0$ yields $\nabla F(x) \to 0$ and $O \to 0$.
LaSalle's invariance principle gives $(x(t), O(t)) \to (\mathcal{X}^\star, 0)$.

Appendix 2. Relation to Variational Free Energy

We briefly compare this to Friston's variational free energy
$F_{\mathrm{var}}(q) = D_{KL}[q(s)\,\|\,p(s|o)] - \ln p(o)$,
whose minimisation yields $q(s) \to p(s|o)$.
In our abstraction, the latent x corresponds to an internal (approximate) posterior, while O reflects a precision-weighted prediction-error direction $\nabla F(x)$. Minimising $F(x)$ aligns the internal state with external evidence; the observer dynamics $\dot{O} = -\gamma O + G\nabla F(x)$ implements a decaying integration of the error signal; the intention feedback KO is bounded by small-gain safety, ensuring a *safe attractor* for free-energy descent.

Appendix 3. Ethical Design Clause

The dynamics are purely mathematical. No personal data are required. The small-gain inequality $\|K\|\,\|G\| < \eta\gamma$ functions as a quantitative "do-no-harm" constraint: intention cannot dominate the natural descent $-\eta\nabla F$. Implementations should avoid centralised logging of internal variables, keep all tuning local, and disclose the chosen gains for accountability.

Appendix 4. Limitations and Future Directions

This framework does not attempt to resolve the *hard problem* of why integrated information feels like something. It provides an *operational and dynamical* description in which consciousness emerges as a measurable variable $O(t)$ that integrates prediction errors over time. The subjective quality (qualia) of experience lies outside the present mathematical scope.

Future work will extend the model to probabilistic and multi-agent contexts by incorporating stochastic free-energy flows and Markov-blanket coupling between agents, enabling empirical tests of conscious-like persistence and attention under controlled simulation.

On scope and phenomenology. This work offers an operational account of consciousness as temporal integration of prediction errors, with formal stability guarantees. It does not attempt to address the qualitative character of experience (qualia). Our framework is compatible with diverse views on phenomenology: it specifies when and how an inferential observable $O(t)$ emerges and persists, while the subjective feel of such states remains outside the present mathematical remit.

References

[1] Friston, K. (2010). *The free-energy principle: A unified brain theory? Nature Reviews Neuroscience*, 11(2), 127–138.

[2] Friston, K., Kilner, J., & Harrison, L. (2006). *A free energy principle for the brain. Journal of Physiology - Paris*, 100(1–3), 70–87.

[3] Parr, T., & Friston, K. J. (2019). *Generalised free energy and active inference. Biological Cybernetics*, 113, 495–513.

[4] Khalil, H. K. (2002). *Nonlinear Systems (3rd ed.)*. Prentice Hall.

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

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

[7] Fleming, S. M., & Dolan, R. J. (2012). *The neural basis of metacognitive ability. Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1338–1349.

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

[9] Lau, H., & Rosenthal, D. (2011). *Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences*, 15(8), 365–373.

Acknowledgements

The author acknowledges the use of AI assistance (ChatGPT GPT-5, OpenAI).