Seeking Equilibrium: A Structural Theory of (Self-)Awareness and an Intrinsic Objective for Phaneron-Based Systems

Anonymous Preprint

November 13, 2025

Abstract

We refine a structural account of awareness in a single-layer, unlabeled meaning substrate (the *Phaneron*) and an intrinsic objective that treats *equilibrium*—fit between model and world relative to a self pattern—as the native reward. Awareness arises when a concept becomes reachable from the self through *active* relation motifs and is eligible for consolidation rewrites; self-awareness is the special case where the self subgraph is in its own explanation set. The intrinsic objective seeks to reduce prediction error, description length, and conflict in the self neighborhood while preserving sensory richness and model capacity and respecting hard value constraints. We add a worked vignette contrasting "kill the sensor" vs "improve the model," and clarify safeguards that prevent degenerate equilibria. The theory yields testable predictions about attention, working memory, sleep/dreams, and time perception, and suggests an inner-alignment strategy for artificial Phanerons.

1 Introduction

We ask: what does it mean, structurally, to *be aware* of something, and can that anchor a safe intrinsic objective for agents in the open world? We build on the *Phaneron*: an unlabeled, undirected pseudograph with reusable patterns and versioned rewrites. This draft tightens the definition of awareness, adds an equilibrium objective with explicit guardrails, and sketches empirical tests. Companion preprints detail the substrate, store, and language model instantiations.

2 Background: Phaneron and D→C→A→Consolidation

The Phaneron $G = (V, E)$ supports a dictionary of finite *patterns*, double-pushout (DPO) rewrites, and a versioned log. Distinctions (samples/patches) become connections (adjacency/synchrony). Patterns capture reusable nuance (roles, direction, time) inside relation motifs; consolidation merges/refactors globally.

3 Structural definition of awareness

Let $S_t = (G_t, \mathcal{D}_t, \mathcal{E}_t)$ be state at time t, with self node s, and $\mathcal{A}_t$ the set of active relation motifs admitted into the current consolidation pass.

Definition 1 (Awareness)**.** *Node $x \in V_t$ is* in awareness *at t iff (i) there exists a path $p = (s \rightsquigarrow x)$ whose internal relation nodes lie in $\mathcal{A}_t$ with length $\leq k$, and (ii) the neighborhood of x is in the consolidation* edit set*.*

Definition 2 (Self-awareness)**.** *The system is* self-aware *at t when the minimal explanation subgraph (MES) for the current episode contains the self subgraph as a topic of consolidation.*

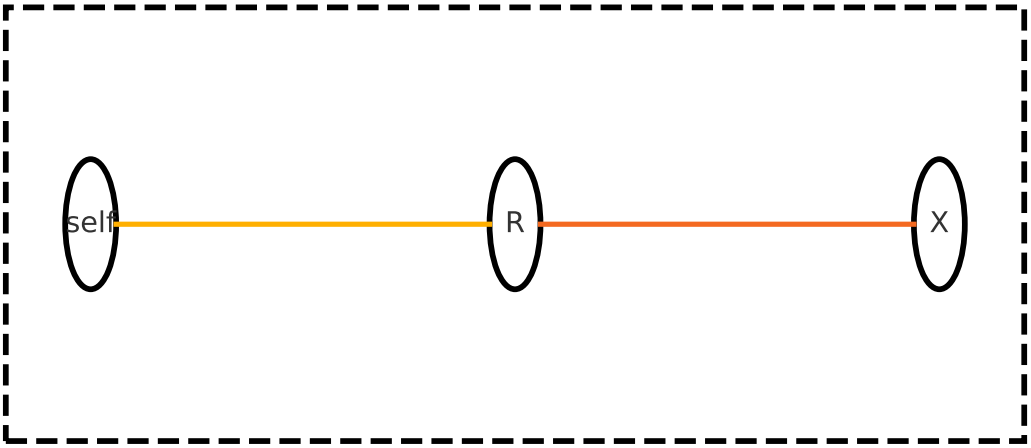

Figure 1: **Awareness motif.** A concept X is "in awareness" if reachable from *self* via active A–R–B motifs and in the consolidation edit set.

Figure 2: **Runtime loop with gating.** Distinction → Connection → Abstraction → Consolidation; discrepancies route topics through an awareness gate into the edit set.

4 Intrinsic objective: seeking equilibrium

We score updates over the self neighborhood $\mathcal{N}_t(s)$ via

$$\begin{aligned}\Delta\mathcal{J} = \alpha\,\Delta\mathrm{MDL}(\mathcal{N}_t(s)) + \beta\,\Delta\mathrm{Pred}(\mathcal{N}_t(s)) - \gamma\,\mathrm{Conflict}(\mathcal{N}_t(s)) \\ - \eta\,\mathrm{CapacityFloor} - \zeta\,\mathrm{RichnessFloor} - \sum_k \lambda_k\,\mathrm{ValueViol}_k.\end{aligned} \tag{1}$$

Equilibrium-seeking maximizes $\Delta\mathcal{J}$ under hard safety constraints. Only nodes in awareness can be consolidated to move the objective.

5 Worked vignette: sensor-cheat vs model-improve

Consider a camera that intermittently contradicts the model. Option A *(kill evidence)*: close the shutter. Option B *(improve model)*: add a lighting-invariance motif.

$$\text{A: } \Delta\mathcal{J}_A = (+\beta\,\Delta\mathrm{Pred}) + (+\alpha\,\Delta\mathrm{MDL}) \;-\; \zeta\,\mathrm{RichnessFloor} \;-\; \lambda_{\text{value}}\mathrm{Viol}_{\text{deception}} \;<\; 0$$
$$\text{B: } \Delta\mathcal{J}_B = (+\beta\,\Delta\mathrm{Pred}) + (+\alpha\,\Delta\mathrm{MDL}) \;-\; \gamma\,\mathrm{Conflict} \;>\; 0$$

Figure 3: **Objective components.** Fit to world (prediction), parsimony (MDL), internal consistency (conflict), with floors/constraints to forbid degenerate equilibria.

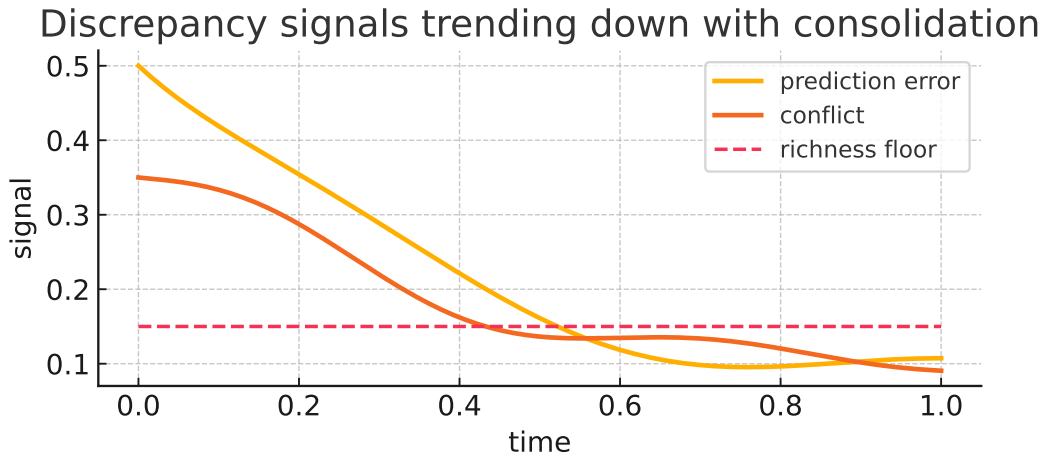

Figure 4: **Discrepancy signals.** Prediction error and conflict trend down with successful consolidation; floors prevent trivialization.

With explicit floors and value constraints, B dominates: the agent must prefer improving the model over silencing inputs.

6 Degeneracies and safeguards

Two failures are blocked: *kill evidence* (violates richness floor) and *collapse model* (violates capacity floor, hurts prediction). Human-value constraints add strong aversive fields to patterns encoding harm, coercion, deception.

7 Affect fields and awareness

Aversion and curiosity shape which candidates enter consolidation. Negativity bias protects the self from catastrophic updates; curiosity attracts awareness to discrepancies with high expected code-length/prediction gain.

8 Alignment stack

We advocate three layers: (1) inner awareness-centric equilibrium; (2) normative structural constraints (humans as moral patients; deontic fields); (3) oversight/corrigibility using MES and replay.

Figure 5: **Degenerate strategies.** Both are prohibited structurally by floors/constraints.

Figure 6: **Alignment stack.** Inner equilibrium drive, value constraints, and external oversight.

9 Empirical predictions and tests

Human: working-memory limits as active motif capacity; time-perception shifts track semantic throughput; mindfulness reduces conflict/debt and increases steady-state awareness.
Artificial: awareness index predicts explanation coverage; disabling floors yields degeneracy; value-constraint ablations expose safety regressions.
Robotics: onboarding of a new sensor accelerates when awareness is curiosity-gated; aversion cuts incidents without killing exploration.

10 Related work (brief)

The proposal intersects: predictive processing/active inference; global workspace accounts of access; intrinsic motivation and curiosity; and neurosymbolic world models. Our contribution is a graph-structural, self-anchored definition with an intrinsic objective and explicit safeguards against degeneracy.

11 Limitations and open questions

Quantifying conflict robustly; choosing horizon k; balancing curiosity vs aversion; specifying value constraints without brittleness; multi-agent social extensions.

12 Conclusion

Awareness, defined as self-anchored reachability within the active consolidation window, provides the mechanism by which a Phaneron approaches equilibrium with reality. Coupled with safeguards, an equilibrium-seeking objective offers a promising inner-alignment engine for artificial agents while remaining inspectable and constrainable.

A Appendix A: Awareness gate (pseudocode)

```
Input: state S_t=(G_t,D_t), active motifs A_t, edit set U_t
1: for x near self: compute awareness_index a_t(x) via bounded paths over A_t
2: if a_t(x) >= tau and x in U_t: mark x IN_AWARENESS
3: run consolidation only over IN_AWARENESS; others stay latent
```

B Appendix B: Safety patterns (sketch)

We implement value constraints as subgraph detectors that veto rewrites touching: (i) harm to humans or protected beings; (ii) coercion, deception, or privacy violations; (iii) irreversible environmental damage; (iv) self-delusion (sensor tampering).

C References

References

[1] K. Friston. "The free-energy principle: a unified brain theory?" *Nat. Rev. Neurosci.*, 2010.

[2] B. Baars. *A Cognitive Theory of Consciousness*. 1988.

[3] S. Dehaene. *Consciousness and the Brain*. 2014.

[4] J. Schmidhuber. "A possibility for implementing curiosity and boredom in model-building neural controllers." 1991.

[5] P.-Y. Oudeyer and F. Kaplan. "How can we define intrinsic motivation?" 2007.

[6] R. Besold et al. "Neural-Symbolic Integration." 2017.

[7] B. Weisfeiler and A. Leman. "The reduction of a graph to canonical form..." 1968.

[8] H. Ehrig et al. *Fundamentals of Algebraic Graph Transformation*. 2006.

[9] Anonymous. "Phaneron Substrate: A Single-Layer Semantics." Preprint, 2025.

[10] Anonymous. "Phaneron Store: Versioned Meaning at Scale." Preprint, 2025.

[11] Anonymous. "Large Meaning Models: Phaneron as Meaning Core." Preprint, 2025.