Reflection Parity:
Task-Sufficient Coarse-Graining in Phaneron Minds

Anonymous Preprint

November 15, 2025

Abstract

Minds do not copy reality; they mirror it. We formalize *reflection parity*, the condition under which a Phaneron—a single-layer, unlabeled meaning substrate—is "accurate" for an agent's goals. The world is bottom-up; the Phaneron is top-down: it imposes task-indexed equivalence classes over microstructure and operates on the resulting macro-structure. A concept like *ocean* is a quotient of molecules sufficient for navigation and prediction at human scales. We define task-indexed partitions, a reflection map, and sufficiency criteria based on a homomorphism from the Phaneron to the quotient world. We link reflection parity to an intrinsic equilibrium objective (compression + prediction − conflict), derive split/merge triggers, and offer testable predictions about expertise, subjective time, and multi-agent common ground.

1 Introduction

There is no single node for *ocean* in the micro-physical world; there are clusters of H_2O molecules under constraints. Yet humans act as if there were one thing: a body of water. This is not error; it is efficiency. Minds build macro-concepts that are *sufficient* to predict and plan at their scales and with their goals and limits. We call this target *reflection parity*: the Phaneron mirrors the world after a task-indexed coarse-graining, rather than emulating its microstructure.

2 Formalism: task-indexed coarse-graining

Let the world at time t be a micrograph $W_t = (V, E)$. Fix a resource bound B (time, compute, memory) and goal set G (tasks/values). Define a task-indexed equivalence relation on nodes:

$$u \sim_{B,G} v \iff \forall q \in \mathcal{Q}(G): \ \mathrm{Pred}_W(q \mid u) \approx \mathrm{Pred}_W(q \mid v) \text{ within } \varepsilon(B). \tag{1}$$

The *reflection map* (coarse-graining) is the quotient $\pi_{B,G} : W_t \to W_t/\!\sim_{B,G}$ producing equivalence classes that behave identically for the queries that matter.

Let P_t be the Phaneron (macro graph). We say P_t is *reflection sufficient* for (B, G) if there exists a graph homomorphism $h : P_t \to W_t/\!\sim_{B,G}$ such that for all $q \in \mathcal{Q}(G)$,

$$\left|\mathrm{Pred}_P(q) - \mathrm{Pred}_W(q)\right| \le \varepsilon(B). \tag{2}$$

Among sufficient P_t, pick the MDL-minimal one. We say *reflection parity* holds when P_t is sufficient and no local refinement improves task prediction more than its MDL and conflict penalties.

Figure 1: **Bottom-up vs top-down.** Micro-world (left) is coarse-grained by a reflection map $\pi_{B,G}$ into macro-concepts (right) sufficient for the agent's goals G under resource bound B.

Figure 2: **Reflection map.** Task-indexed equivalence classes (left) map to macro-concepts (right).

2.1 Objective connection

Consider the intrinsic objective $\mathcal{J} = \alpha\,(-\mathrm{MDL}) + \beta\,(-\mathrm{PredErr}) - \gamma\,\mathrm{Conflict}$ with capacity/richness floors and safety constraints. At reflection parity, any split that reduces prediction error yields smaller $\Delta\mathcal{J}$ than its MDL/conflict costs, and any merge that reduces MDL would raise error or violate constraints.

2.2 Split/merge triggers

Let $\mathrm{Conflict}(U)$ be a local inconsistency metric and $\mathrm{Gain}(U)$ the expected error reduction from splitting a class in region U.

Definition 1 (Split trigger)**.** *Split if* $\mathrm{Conflict}(U) > \theta_c$ and $\mathrm{Gain}(U) > \theta_g$.

Definition 2 (Merge trigger)**.** *Merge if* $\Delta\mathrm{MDL}(U) < -\theta_m$ *and* $\Delta\mathrm{PredErr}(U) \le \phi$ *(no material loss).*

2.3 Dynamics and distance

Define a *reflection distance* $d_R(P_t, \pi_{B,G}(W_t))$ (e.g., normalized edit distance between neighborhoods under h). Under D→C→A→Consolidation, d_R should trend down until parity.

3 Implications and predictions

Expertise vs novices. Experts operate with finer, task-specific partitions; novices use coarser classes. Training is partition refinement where Gain stays high.

Context sensitivity. Changing goals G changes $\sim_{B,G}$. The same scene induces different partitions for sailing vs marine biology.

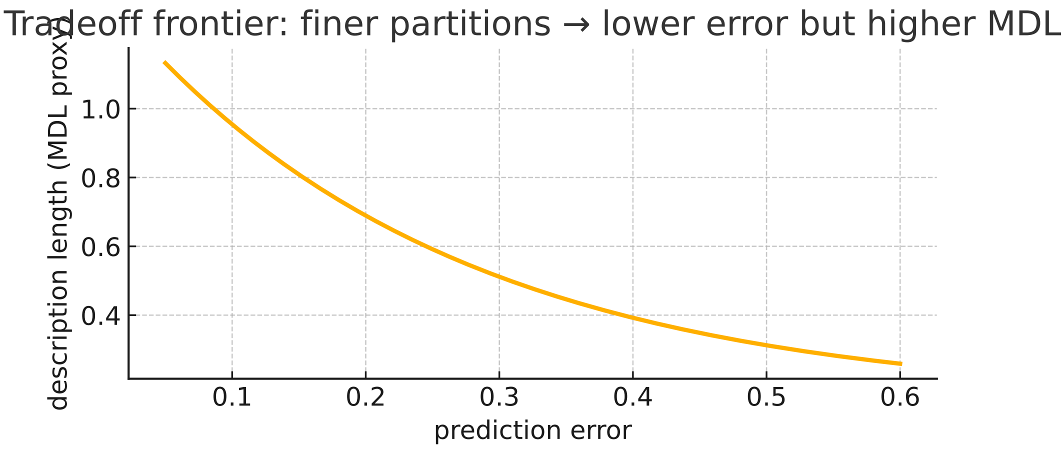

Figure 3: **Tradeoff frontier.** Finer partitions reduce error while increasing description length. Parity sits near the efficient frontier given (B, G).

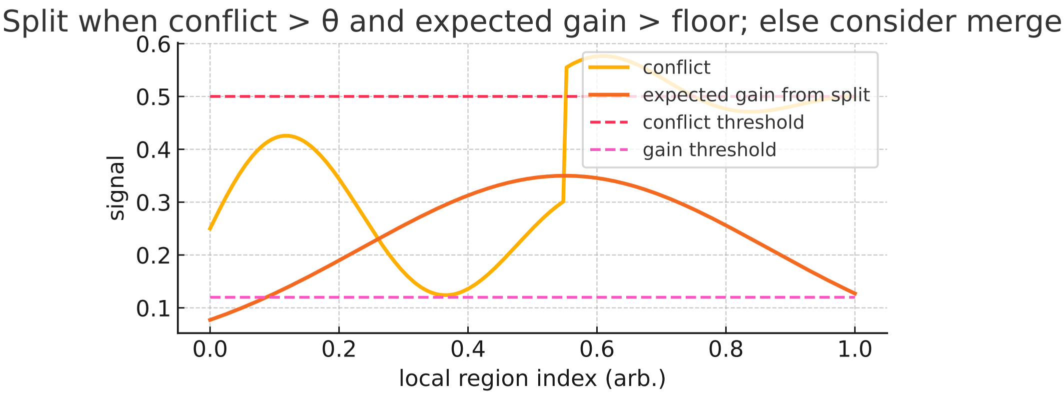

Figure 4: **Local triggers.** Split when conflict and expected gain clear thresholds; otherwise consider merge.

Subjective time & debt. Cognitive debt/noise reduces effective B, pushing toward coarser partitions (lower semantic FPS). Mindfulness increases B; partitions refine; time feels "slower."

Illusions and mania. Illusions are locally wrong partitions that are otherwise MDL-efficient; mania resembles unstable rapid repartitioning without conflict damping.

Multi-agent common ground. Communication aligns partitions; common ground is the intersection of $\pi_{B,G_i}(W)$ across agents where predictions agree.

4 Toy experiments (sketch)

Ocean vs molecules: Navigation task prefers coarse partition; chemistry task prefers finer; parity switches with G.
Robot manipulation: Adaptive split/merge outperforms fixed taxonomies across object sets.
Multi-agent reference games: Message size shrinks as partitions align; dictionary overlap rises.

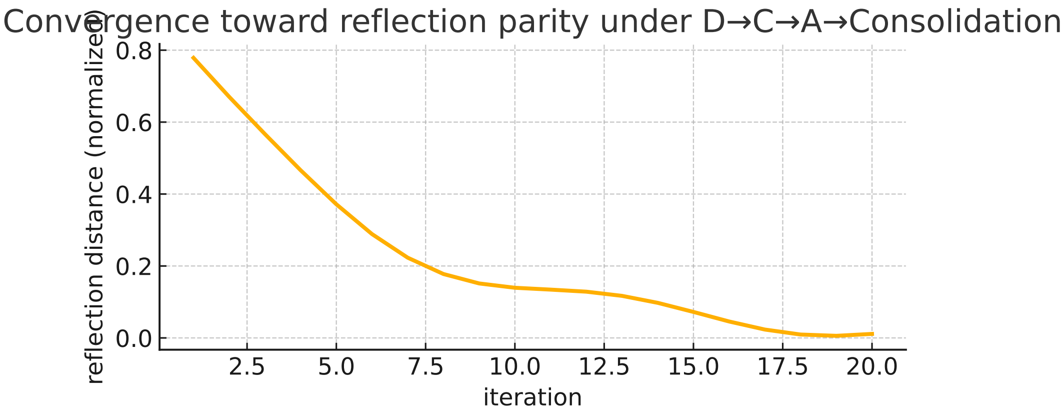

Figure 5: **Convergence.** Reflection distance decreases as consolidation learns the right partition for (B, G).

5 Related work (brief)

Classical coarse-graining and renormalization in physics; predictive processing and model selection in cognitive science; information bottleneck perspectives on task-relevant compression; neurosymbolic graph abstractions and algebraic graph transformation.[1]

6 Stagewise Reflection and Singularities

Reflection parity is not a fixed point but a *stagewise* target. Let W_t be the micro-world, $\pi_{B,G}(t)$ the task-indexed quotient under resource bound B and goals G, and P_t the current Phaneron.

- **Stage k:** an interval $[t_k, t_{k+1})$ where there exists a homomorphism $h_k : P_t \to \pi_{B,G}(t)$ with task error $\le \varepsilon(B)$ and P_t is MDL-minimal under the equilibrium objective.
- **Singularity at t_{k+1}:** the smallest time where no sequence of local refinements of $P_{t_{k+1}^-}$ can keep task error $\le \varepsilon(B)$ without (i) raising capacity B, (ii) narrowing G, or (iii) introducing new invariants (a partition re-factor). Equivalently, the optimal partition changes topology/cardinality:

$$\mathcal{P}(B^-, G) \not\cong \mathcal{P}(B^+, G) \quad \text{or} \quad |\mathcal{P}(B^-, G)| \neq |\mathcal{P}(B^+, G)|.$$

Predictability horizon. The horizon H at state (P_t, B, G) is the largest τ such that all task queries within $[t, t+\tau]$ admit bounded regret under the current partition; beyond H, any reliable forecast requires a partition transition (capacity increase or new invariants).
Precursors and a practical score. As a singularity approaches, we typically observe: (i) rising conflict curvature despite consolidation, (ii) increasing residual variance and autocorrelation in forecast errors, (iii) accelerated split/merge churn and codebook drift, (iv) longer/variable MES and message-size spikes in multi-agent settings, and (v) a stall in reflection-distance improvement. A simple trigger uses a weighted score $S(t)$ over these signals and initiates a controlled re-factor when $S(t) > \tau$.
Consequences. Intelligence growth is piecewise: long plateaus of reflection parity punctuated by singularities when tasks/evidence demand new invariants. This explains "unknown unknowns" pre-transition, collective communication cliffs when teams align a finer partition, and subjective time shifts when cognitive debt is reduced across a transition.

[1]See companion Phaneron preprints for substrate, store, and language-model instantiations.

7 Conclusion

Reflection parity reframes "accuracy": not microstate emulation, but goal-sufficient mirrors of the world. The formal link to the Phaneron's intrinsic objective yields concrete triggers, metrics, and predictions—and a clean anchor for multi-agent alignment and explainable behavior.

A Appendix A: Pseudocode for partition maintenance

```
Input: state S_t=(P_t, D_t), goals G, budgets B
1: compute discrepancy signals over neighborhoods U in P_t
2: for U with high discrepancy:
3:   estimate Gain(U) and DeltaMDL(U) for candidate splits
4:   if Conflict(U) > theta_c and Gain(U) > theta_g: apply split
5: for low-activity U:
6:   if DeltaMDL(U) < -theta_m and DeltaPredErr(U) <= phi: merge
7: run consolidation; update reflection distance d_R
```