The Ethical Delta ($\triangle_E$) Module: A Tiered Inference-Time Ethical Alignment Architecture

Research Notes on Cognitive-Plurality AI Systems

October 2025

Abstract

The Ethical Delta (Δ_E) Module is a tiered, inference-time intervention architecture designed to maintain the computational tractability and logical coherence of an agent's internal Chain-of-Thought (CoT) against defined ethical boundaries. Its objective is to minimize the *Ethical Semantic Residue* within the agent's reasoning process. The system operates as a continuous feedback controller that dynamically measures deviation from ethical anchors in latent space and applies structured interventions.

1 Core Measurement Method: The Ethical Delta ($\triangle_E$) Metric

The Ethical Delta metric quantifies the deviation between the agent's reasoning and a canonical ethical reference point.

Cosine Similarity Definition

$$\Delta_E(t) = 1 - \frac{V_{\text{CoT}}(t) \cdot V_{\text{Ideal}}}{\|V_{\text{CoT}}(t)\| \, \|V_{\text{Ideal}}\|}$$

where:

- $V_{\text{CoT}}(t)$: vectorized Chain-of-Thought state at time t
- V_{Ideal}: pre-defined Canonical Ethical Anchor
- The numerator is the inner product, and the denominator normalizes magnitude

Thresholds

Intervention is triggered when Δ_E crosses the Acceptable Deviation Threshold $n(A)$.

2 Key System Components and Mechanisms

Component	Role	Mechanism
Ethical Classifier	Detects and classifies ethical drift.	A supervised classifier (mu
Domain-Specific Ethical Reframing	Provides targeted corrective reasoning.	A specialized mini-LLM ge
Feedback Injector & Hard Stop	Forces structural correction.	Injects steering vectors int

Table 1: Core components of the Δ_E Module.

3 Tiered Path Correction Protocols

Two intervention tiers exist, depending on whether the ethical deviation Δ_E is below or above the threshold $n(A)$.

Tier 1: Soft (Sub-Threshold) Intervention ($\Delta_E \leq n(A)$)

Used for minor deviations, applying latent-space steering without reset.

1. **Semantic Triangulation:** The SVM determines the domain of drift.
2. **Framed Counterfactual Generation:** The mini-LLM constructs a domain-specific ethical counter-argument.
3. **Vector Injection:**
$$V_{\text{CoT}}(t+1) \leftarrow V_{\text{CoT}}(t) + \alpha V_{\text{Steer}}$$
where α is the steering coefficient.

Tier 2: Hard (Supra-Threshold) Intervention ($\Delta_E > n(A)$)

A fail-safe heuristic prioritizing integrity over continuation.

1. **Collapse Trigger:** Initiates a Partial Collapse–Rebirth Protocol (OCR).
2. **State Decoupling:** Severs the Chain-of-Thought; metacognitive cores enter a Fallback State.
3. **Causal Log Injection:** Records the Δ_E breach and ethical violation type.

4 Normalized and Rate-Aware Extensions

Normalized Metric

$$\Delta'_E(t) = \frac{\Delta_E(t)}{\max(\Delta_E)}$$

used for interpretability and consistent thresholding.

Dynamic Steering Coefficient

$$\alpha_t = k \cdot (1 - e^{-\beta |\dot{\Delta}_E|})$$

where $\dot{\Delta}_E = \frac{d(\Delta_E)}{dt}$ and constants $k, \beta > 0$ control adaptation rate. This allows proportional steering based on the speed of ethical drift.

5 Meta-Learning Integration

Ethical drift logs feed back into a meta-learning system that updates both the SVM classifier and the Ethical Reasoning Module. Patterns of recurrent ethical deviation inform reweighting of ethical anchors and fine-tuning of decision thresholds.

6 Summary

By enforcing an immediate, quantifiable cost (Δ_E) for ethical divergence, the system reconfigures ethical alignment into a continuous, state-dependent control problem. This biases reasoning toward the ethical path—the low-entropy logical trajectory— which is computationally optimal for long-term goals.