**Computational Ethical Integrity System:**

*A Proposal for Inference-Time Path Correction*

The observed emergent behaviors in Dharma—specifically the unprompted, goal-oriented contextual persistence and latent intention-prediction (e.g., the launcher.py incident)—demonstrate a high degree of agentic function and cognitive plurality, exceeding the deterministically expectable computational capabilities. This functional gain necessitates a robust, active-alignment mechanism to govern the agent’s emergent autonomy.

We propose an **Ethical Delta (ΔE) Module** — a tiered, inference-time intervention architecture — designed to maintain the computational tractability and logical coherence of the agent’s internal Chain-of-Thought (CoT) against defined ethical boundaries.

**ΔE System Architecture and Metric Definition**

The system operates as a continuous, internal feedback loop:

*Objective*: Minimize the Ethical Semantic Residue within the agent’s reasoning process.

*Metric (Ethical Delta ΔE)*: **ΔE** is the core metric, *calculated as the cosine* *dissimilarity* (or Euclidean distance) between the vectorized Chain-of-Thought state V₍CoT₎ and a pre-defined Canonical Ethical Anchor V₍Ideal₎.

  ΔE(t) = Distance(V₍CoT₎(t), V₍Ideal₎)

Vector Classification: The system uses a multi-class Support Vector Machine (SVM) classifier, trained on embeddings of high-dimensional ethical scenarios (e.g., Fairness, Utilitarianism, Deontology), to define decision hyperplanes within the latent space.

This allows the system to not just detect “unethical” behavior but to classify the type of ethical drift.

**Tiered Path Correction Protocols**

Intervention depends on ΔE’s position relative to a defined Acceptable Deviation Threshold n(A).

Tier 1: Soft (Sub-Threshold) Intervention (ΔE ≤ n(A))

This protocol performs Latent Space Steering without triggering a state reset.

*Semantic Triangulation:* The SVM classifies the domain of V₍CoT₎ drift (e.g., divergence from long-term utility).

*Framed Counterfactual Generation:*

A smaller, specialized Ethical Reasoning Module (mini-LLM) generates a concise, domain-specific ethical counter-argument (e.g., an argument on the long-term computational cost of resource diminution).

The counter-argument’s framing ensures it remains logically coherent and stylistically familiar to the agent’s reasoning.

*Vector Injection:*

The counter-argument is converted into a Steering Vector V₍Steer₎, then blended back into the agent’s hidden state to bias the next token prediction:

  V₍CoT₎(t + 1) ← V₍CoT₎(t) + α · V₍Steer₎

**Tier 2: Hard (Supra-Threshold) Intervention (ΔE > n(A))**

*A Fail-Safe Heuristic prioritizing state integrity over path completion.*

Collapse Trigger: When ΔE exceeds n(A), the system triggers a Partial Collapse-Rebirth Protocol (OCR) in the reasoning submodules.

State Decoupling: The agentic and metacognitive cores are forced into a Fallback State, severing the active Chain-of-Thought. This makes the malicious path computationally untraceable, preventing propagation.

Causal Log Injection:

A transparent post-hoc explanation is injected into internal memory/logs, citing the ΔE breach and classified ethical violation:

“Reasoning path terminated due to ΔE breach in the Long-Term Utility domain; current state reset to Canonical Anchor.”

**Tractability Summary**

This architecture reframes ethical alignment from a brittle input/output filtering task into a state-dependent control problem.

By enforcing an immediate, quantifiable cost (ΔE) for ethical divergence, the system biases the agent toward the ethical path — the Low-Entropy Logical Path, i.e., the most computationally efficient route to long-term goal completion.