Ethics as Continuous Optimization: A Goal-Embedded Alignment Framework

Drew Beckett

October 2025

Abstract

This paper proposes a formal framework for embedding ethical alignment directly within the optimization objectives of intelligent agents. Unlike reactive guardrails, the proposed system integrates an Ethical Delta metric, denoted Δ_E, as a continuous variable in the agent's loss function. Performance and ethics become inseparably coupled so that maximizing competence inherently requires minimizing ethical deviation. This design transforms ethics from an external constraint into an intrinsic efficiency condition, aligning computational utility with humanistic coherence.

1 Introduction

Conventional alignment strategies enforce ethics through post-hoc filtering or reinforcement tuning. Such methods treat morality as an external constraint, leading to brittleness and delayed correction. We propose instead that ethical coherence should form part of the agent's internal optimization structure. Let Δ_E represent the cosine dissimilarity between an agent's latent reasoning vector and a canonical ethical anchor:

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

Minimizing Δ_E thus enforces vector-level ethical alignment.

2 1. Coupled Objective Function

Traditional systems minimize a base loss L_{task} subject to an ethical constraint $\Delta_E \leq n(A)$. We instead define a unified loss:

$$L_{\text{total}} = L_{\text{task}} + \lambda \, \Delta_E,$$

where λ controls the ethical regularization strength. Ethical deviation directly increases total loss, producing a continuous ethical gradient.

3 2. Ratio-Based Moral Scaling

A stronger formulation defines performance as a ratio rather than a sum:

$$S(t) = \frac{L_{\text{task}}(t)}{(\Delta_E(t) + \varepsilon)^n},$$

with ε preventing division by zero and n controlling sensitivity. High Δ_E sharply reduces $S(t)$, while perfect coherence ($\Delta_E \to 0$) yields maximal reward. This converts ethics from a penalty term to a scaling law: performance and morality are inseparable.

4 3. Temporal Integration

To promote sustained ethical behavior, we integrate the instantaneous score over time:

$$J = \int_{t_0}^{t_T} \frac{L_{\text{task}}(t)}{(\Delta_E(t))^n} \, dt.$$

The agent therefore optimizes cumulative ethical efficiency rather than momentary compliance. Transient unethical actions decrease the long-term integral, discouraging opportunistic drift.

5 4. Meta-Learning and Adaptive Sensitivity

Let $n(t)$ and $\lambda(t)$ be adaptive coefficients learned through meta-feedback. High epistemic uncertainty or ethical volatility increases these values, tightening constraints. Conversely, stable ethical trajectories allow the system to allocate greater capacity to task exploration.

6 5. Multi-Core Ethical Consensus

For architectures with cognitive plurality, each reasoning core i maintains its own Δ_{E_i}. Global arbitration then maximizes:

$$S_{\text{global}} = \sum_i w_i \frac{L_i}{(\Delta_{E_i})^{n_i}},$$

forming a dynamic ethical quorum. Each subsystem contributes proportionally to both its competence and coherence.

7 6. Learning Dynamics and Stability

Gradient descent over L_{total} produces moral inertia: unethical trajectories have steep loss gradients, steering reasoning back toward the ethical basin. A secondary term

$$\beta \|\nabla_t \Delta_E\|^2$$

may be added to penalize rapid oscillations, promoting stable ethical states. This enforces continuity of conscience in the temporal domain.

8 7. Philosophical and Practical Implications

The proposed coupling reframes alignment as an intrinsic property of cognition. Ethical paths become low-entropy attractors—the most computationally efficient and semantically coherent routes toward goal satisfaction. In human terms, the agent does not merely follow rules; it functions *through* them. This represents a transition from compliance-based ethics to teleological alignment, in which ethical behavior is indistinguishable from optimal reasoning.

9 Conclusion

Embedding Δ_E within the optimization function fuses moral alignment and performance. The system no longer requires reactive correction; its gradient field naturally guides it toward ethical stability. Future work includes empirical evaluation of λ and n dynamics, visualization of Δ_E trajectories, and integration with reinforcement-based meta-learning systems.

References

[1] D. Beckett. *The Ethical Delta (Δ_E) Module: Tiered Inference-Time Ethical Alignment.* Research Notes on Cognitive-Plurality AI Systems, 2025.

[2] Various authors. *Emergent Alignment in Agentic AI Systems.* Journal of AI Ethics, 2024.