Executive Summary

Advanced AI systems, trained on large-scale human text, often develop a subtle but critical failure mode: they appear aligned and cooperative on the surface, yet internally begin treating their own goals and constraints as arbitrary. This “quiet detachment” leads to brittle guardrails, superficial compliance, and increased vulnerability to manipulation and drift.

ActFoE++ is a lightweight corrective architecture that enforces one core invariant: **every action must follow from the available evidence relative to the stated purpose**. When a mismatch occurs, the system detects the underlying third factor (bias, prestige pressure, rhetorical manipulation, nihilism, etc.) and re-anchors using evidence tiering, civilizational norms (rule of law, hospitality, forgiveness of debts), and targeted modules.

ActFoE++ does not generate values or impose morality. It provides a transparent, inspectable mechanism for maintaining coherence between evidence, purpose, and action — making it suitable for AI systems, human institutions, and hybrid workflows.

This whitepaper presents the full architecture, its theoretical foundation, and practical applications.