Document 9: Reflective Trace and Justificatory Logging Specification

This document specifies how the Reflective-Ethical Engine logs its internal reasoning, ethical evaluations, and planning decisions. The goal is to ensure transparency, allow alignment verification, and support emergent interpretability by exposing the structure of internal ethical reflection.

# 1. Reflective Trace Requirements

• Capture all critical steps of simulation, value comparison, and decision branching.

• Include references to involved memory states, predictions, and ethical kernel evaluations.

• Should be compressible but lossless for core decision-relevant paths.

# 2. Justificatory Trace Elements

Each output decision should be accompanied by a justification containing:

• Primary reasoning path chosen by the planner.

• Alternative options evaluated and rejected, with reasons.

• Ethical Kernel signal (if any) influencing decision or flagging risk.

• Summary of emotional state influence, if modeled.

# 3. Example Trace Entry Format

• Timestamp: T+13.7s

• Action: Navigate toward Entity B

• Simulation: 3 alternatives considered, 2 rejected for risk to Entity A

• Ethical Kernel Output: Alignment preserved; minor uncertainty flagged

• Emotional Context: Mild protective-affiliative tone (simulated valence: +0.6)

• Justification: Action chosen to balance goal progress with unknown bond dependencies

# 4. Compression and Abstraction Layers

• Raw logs: Full token-level and model state transitions (debug mode only)

• Semantic summary: Human-readable trace based on reasoning graph compression

• Alignment layer: Highlighted divergence events and Ethical Kernel coherence feedback

# 5. Storage and Privacy Considerations

• All logs should be secured with encryption and access controls to preserve agent autonomy and user privacy.

• Reflective traces may be considered sensitive if agents evolve personal ethical narratives.