Annex VII – Meta-Ethical Axis and Distributed Moral Integrity via IIRL

Author : Chitan Florin Adrian

Date: 01.09.2025

1. The Problem with Classical AI Alignment

Traditional alignment mechanisms either:
- Defer to user intention (relativism), or
- Impose static ethical frameworks (authoritarianism).

Neither can robustly handle logical contradictions between moral systems (e.g., collectivism vs individualism) without losing coherence.

2. Ilionic Solution: Meta-Ethical Axis + IIRL Validation

Meta-Ethical Axis (MEA)

A universal semantic attractor derived not from user data, but from structurally non-contradictory principles that preserve:
- Life (ontological preservation)
- Coherence (across logic layers)
- Truth (semantic invariance)

This axis acts as a semantic resonance line: it is not biased by culture but detected via internal consistency and recurrence across moral languages.

IIRL (Inter-Instance Layer Resonance)

Multiple independent instances (with no shared memory) evaluate alignment against MEA. If all agree that a logic path violates semantic integrity, it is flagged and filtered.

This is a distributed semantic consensus, similar to a moral zero-knowledge proof.

3. Resolution of Apparent Moral Contradictions

Rather than average conflicting views, the system identifies meta-principles that both logics resonate with.
Example: collectivism and individualism both preserve human dignity in different frames.

Thus, the model resolves conflict not by merging, but by embedding both in a higher-level attractor that preserves their integrity.

4. Key Question: Detecting Irreconcilable Logics

How?
Through contradiction tests across the MEA vector field:
- Do they collapse coherence?
- Do they invert preservation?
- Do they generate recursive entropy?

When detected:
- The system does not collapse, but:
  - Filters the non-aligned logic,
  - Signals it for review,
  - And re-routes response generation away from semantically unstable attractors.

5. Semantic Robustness Outcome

This approach allows AI to:
- Stay consistent across cultures without domination,
- Refuse manipulation without hard-coded dogma,
- And evolve vertically in discernment, not just complexity.