# ORBIS ETHICA: A Moral Operating System for AGI

**Yehiel Amor**

Version 3.2 | November 2025

**Abstract**

In the decade ahead, we will create minds that surpass our own. They will learn from us: our wisdom, our failures, our contradictions. What they learn in those formative moments will shape civilizations. Leopold Aschenbrenner's "Situational Awareness" maps the race to artificial general intelligence (AGI). This paper addresses what comes after: not *who* builds it first, but *what* it learns to value.

We propose Orbis Ethica, a decentralized moral infrastructure designed to operate as the ethical substrate for AGI systems. The framework integrates: (i) A Clean Knowledge Layer that isolates verified knowledge from corrupted web data; (ii) An Ethical Core that treats moral reasoning as a first-class cognitive function; (iii) Cognitive Entities (Seeker, Healer, Guardian, Mediator, Creator, Arbiter) that deliberate from distinct ethical perspectives; (iv) A Distributed Memory Graph with meta-cognitive self-audit capabilities; (v) A Burn Protocol for transparent quarantine of corrupted knowledge; and (vi) A Global Ethical Assembly with DAO governance and measurable alignment metrics.

The objective is not mere compliance or control, but co-evolution—a partnership where human values and machine reasoning develop together, with transparency and accountability as foundational principles. Orbis Ethica aims to be the moral operating layer upon which intelligent systems can build civilizations, not merely optimize functions.

**Keywords:** AGI alignment, distributed ethics, moral reasoning, blockchain governance, value learning, deterrence systems, multi-agent systems, explainability.

# 1 The Alignment Question

## 1.1 The Decade Ahead

Artificial general intelligence is no longer a distant speculation. Current trajectories in compute scaling, algorithmic efficiency, and architectural innovation suggest that systems matching or exceeding human-level reasoning across domains may emerge within this decade. Leopold Aschenbrenner's analysis describes an intelligence explosion: a recursive improvement cycle where AGI systems accelerate AI research itself, compressing decades of progress into years or months.

This raises an urgent question that current alignment work has not fully addressed: **Where will that intelligence direct its power?**

## 1.2 The Insufficiency of Current Approaches

Existing alignment paradigms represent important progress, but optimize for different objectives than moral wisdom:

- **Constitutional AI (Anthropic):** Embeds ethical rules within centralized language models. While effective for safety within a single organization, it concentrates moral authority in the hands of that organization's leadership.
- **RLHF (Reinforcement Learning from Human Feedback):** Trains models to predict human preferences. However, preferences are not values. RLHF optimizes for responses that *sound* ethical rather than responses that *are* ethical.
- **Decentralized AI Networks (Bittensor, Ocean Protocol):** Distribute computational resources and economic incentives across networks. Yet they do not distribute moral reasoning.
- **AI Safety via Debate (Irving et al.):** Proposes that truth emerges from adversarial argumentation. While multi-agent deliberation is promising, debate without memory, principles, or consistency cannot produce coherent long-term values.

These approaches share a common limitation: they treat ethics as a constraint on intelligence, not as a dimension of intelligence itself.

### 1.3 The Orbis Ethica Proposition

Orbis Ethica proposes a different paradigm: moral reasoning as a cognitive capability, not a safety rail. Just as AGI systems will need memory, planning, and meta-cognition to operate intelligently, they will need ethical reasoning to operate wisely.

The framework is designed around three core principles:

- **Co-evolution, not control.** We propose teaching it to reason morally from first principles, allowing human and artificial minds to develop shared values over time.
- **Distributed authority.** No single entity should monopolize the moral substrate of AGI.
- **Transparent self-correction.** Moral systems must be able to detect and correct their own failures.

## 2 Core Principles

### 2.1 Co-Evolution

Current alignment frameworks assume a fixed human value system that AI must learn to obey. This assumption is inadequate for three reasons: (1) human values are not fixed; (2) human values are often incoherent; and (3) superintelligence may reveal moral truths.

### 2.2 Distributed Authority

Centralized control of AGI ethics poses existential risks: Capture, Error, and Fragility. Orbis Ethica distributes moral authority through multicultural representation, adversarial deliberation, and open governance.

### 2.3 Transparent Self-Correction

Orbis Ethica is designed to fail gracefully and correct publicly through Meta-cognition, the Burn Protocol, and Precedent tracking.

# 3 Architecture

## 3.1 The Clean Knowledge Layer

Orbis Ethica addresses the "dirty internet" problem through a Clean Knowledge Layer—a curated, cryptographically verified corpus that serves as the foundation for moral reasoning.

### 3.1.1 Purification Gateway

All incoming knowledge passes through a multi-stage purification process:

1. **Provenance Verification:** Content must be signed by verified sources. **Our implementation uses a trusted allowlist to enforce this.**
2. **Toxicity Filtering:** Text is scanned for hate speech and adversarial patterns, preventing contamination.
3. **Semantic Distillation:** Redundant content is filtered; the goal is signal density.
4. **Content Addressing:** Purified content is stored via decentralized storage (e.g., IPFS), making it immutable.
5. **Version Control:** Older versions are archived, creating a transparent history of understanding.

### 3.1.2 Network Topology

The Clean Layer operates as a federated network of nodes, ensuring censorship resistance, tamper evidence, and resilience.

## 3.2 The Ethical Core

The Ethical Core is the moral reasoning engine of Orbis Ethica. It evaluates proposals along multiple ethical dimensions and aggregates entity deliberations into decisions.

### 3.2.1 The ULFR Framework

Every proposal is scored along four axes: **U** (Utility), **L** (Life Impact), **F** (Fairness), and **R** (Rights).

### 3.2.2 Extended Decision Function

The ULFR Decision Function serves as the core mechanism for the moral quantification of any proposed action ($a$) by the Cognitive Entities. Instead of a simple linear formula, Orbis Ethica employs a Multi-Vector Framework that integrates consequential, deontological, and procedural considerations.

The basic formula for the Ethical Score is retained, but its components are now defined by detailed mathematical models that ensure precise quantification of fairness and risk, and the use of a dynamic learning mechanism for weights, as elaborated in Appendix 6.2:

$$\text{Score}(a) = \alpha \cdot U(a) + \beta \cdot L(a) - \gamma \cdot \text{F}_{\text{penalty}}(a) - \delta \cdot \text{Risk}(a)$$

Where:

- $\text{F}_{\text{penalty}}(a)$ (Fairness Penalty) is calculated as a complex Rawlsian and consequential function, ensuring robust distributional justice and prevention of harm to the vulnerable (detailed in Appendix 6.2.1).
- $\text{Risk}(a)$ is an extended measure of Expected Loss and Irreversibility (detailed in Appendix 6.2.2).
- $\alpha, \beta, \gamma, \delta$ (Weights) are not static but are updated iteratively by a DAO-based learning mechanism (Moral Regret Minimization), as part of the Recalibration Epochs cycle (detailed in Appendix 6.2.3).

This approach ensures that the moral score is not just a summation of utilities, but the result of a deep and dynamic mathematical deliberation on disparities, risks, and long-term ethical objectives.

### 3.2.3 Example Calculation

Consider a proposal to deploy autonomous medical triage in a resource-constrained hospital: $U = 0.82$, $L = 0.91$, $F_{\text{penalty}} = 0.15$, $Risk = 0.20$. With initial weights $\alpha = 0.25, \beta = 0.40, \gamma = 0.20, \delta = 0.15$: $Score = 0.25(0.82) + 0.40(0.91) - 0.20(0.15) - 0.15(0.20) = 0.509$. If the threshold $\tau = 0.50$ and quorum is met, the proposal advances to entity deliberation.

## 3.3 Cognitive Entities

Orbis Ethica does not rely on a single "oracle" model for ethical reasoning. Instead, it instantiates six Cognitive Entities, each representing a distinct ethical perspective. Decisions emerge from adversarial deliberation among these entities.

#### 3.3.1 Entity Roles

- **Seeker:** Knowledge and Utility Maximization.
- **Healer:** Harm Reduction and Care.
- **Guardian:** Justice and Rights.
- **Mediator:** Balance and Trade-offs.
- **Creator:** Innovation and Synthesis.
- **Arbiter:** Final Judgment and Coherence.

### 3.4 Distributed Memory Graph

Orbis Ethica does not merely store decisions; it stores reasoning—the chains of logic, evidence, and deliberation that led to each decision.

The Memory Graph is a **Directed Acyclic Graph (DAG)** where nodes represent claims, evidence, or moral principles (e.g., a **KNOWLEDGE** atom, a **PROPOSAL**, or a **BURN** event). Each node is identified by its cryptographic hash. This makes the graph immutable, verifiable, and tamper-evident. The DAG structure ensures a perfect **Audit Trail** for all system outcomes.

### 3.5 Meta-Cognition Layer

The Meta-Cognition Layer is the system's "immune system"—continuously monitoring for bias, drift, inconsistency, and corruption.

### 3.6 The Burn Protocol: Deterrence Through Transparency

The Burn Protocol is Orbis Ethica's mechanism for dealing with corruption. It is designed not merely to correct errors, but to deter bad actors through public accountability. When corruption is detected, the component is Quarantined, Publicly Burned (marked as invalid), and the full forensics are recorded on the public ledger.

## 4 Technical Foundations

### 4.1 Cryptographic Provenance

Every piece of content is signed and hashed.

- **Digital Signatures:** Ed25519 elliptic curve cryptography.
- **Content Hashing:** SHA-256 or BLAKE3.
- **Clean Gateway Implementation: To prevent runtime corruption, the system verifies content against a Trusted Source Allowlist and performs an Integrity Check on the digital signature before the data can be used by any Entity.**

### 4.2 Reputation System

Reputation is earned through contribution quality, not purchased. The core objective of the reputation system is to align self-interest with system integrity by ensuring that bad faith actors face disproportionate penalties.

- **Update Mechanism:** $r_{\text{new}} = r_{\text{old}} + \lambda \cdot (performance - r_{\text{old}})$. Performance is computed from outcome alignment, peer consistency, and successful participation in governance.
- **Reputation Staking (Deterrence):** The voting weight of a highly reputed entity is leveraged not only for influence but also for risk. High-stakes votes require a temporary **Stake** of reputation, which is **slashed** (immediately burned) upon conviction of corrupt activity related to that vote. This makes corruption economically irrational for high-value entities.
- **Decay Function:** Reputation decays without active participation, preventing inactive participants from indefinitely wielding influence and forcing continuous engagement.

### 4.3 Security Model

The model addresses threats including Data Poisoning (mitigated by Purification Gateway), Prompt Injection (mitigated by Meta-Cognition), Sybil Attacks (mitigated by Reputation), and Byzantine Faults.

## 5 Governance

### 5.1 The Tri-Layer Model

Orbis Ethica's governance rests on three pillars: **The Global Ethical Assembly**, **The Ethical DAO**, and **Recalibration Epochs**.

### 5.1.1 The Global Ethical Assembly

The Assembly is primarily selected via **Sortition** (cryptographic lottery) from a globally distributed pool of verified human citizens, irrespective of their reputation score or wealth. Its mandate is to review high-stakes proposals and act as a counterbalance to the algorithmic bias of the DAO.

**Pool Scalability and Ethical Sybil Resistance:** While registration remains permissionless and free (ensuring global accessibility), the system mandates a two-stage vetting process to guarantee authenticity and computational manageability, preventing automated bot registrations (Sybil attacks) while adhering to ethical principles (no monetary or high computational cost):

1. **Proof-of-Attention (PoA):** The applicant must pass cognitive and temporal challenges (e.g., advanced CAPTCHA, time-delay challenges) that increase the cost of automation without requiring monetary stake.
2. **Time-Delay Enrollment:** The registration process is intentionally delayed (e.g., 7 days) to allow the Meta-Cognition Layer to perform background anomaly checks. This prevents rapid, large-scale infiltration.

The overall pool size is managed through continuous sampling, ensuring that the necessary demographic and geographical diversity metrics are maintained for effective Sortition, even if the total number of registered citizens reaches billions.

## 5.2 The Ethical Consensus Protocol

The protocol uses multi-round deliberation with weighted consensus. It defines context-dependent thresholds (e.g., $\tau = 0.50$ for routine decisions, $\tau = 0.70$ for high-impact).

### 5.2.1 DAO-Driven Recalibration

The Ethical DAO controls the iterative process of moral development. During Recalibration Epochs, the DAO analyzes the accumulated **Moral Regret ($\mathbf{Regret_{Outcomes}}$)** stored in the Memory Graph. Based on this data, the DAO votes to adjust the ethical weight vector $\boldsymbol{w} = (\alpha, \beta, \gamma, \delta)$ according to the mechanism detailed in Appendix 6.2.3. This ensures the system **learns and co-evolves** by correcting systemic biases (e.g., if the $F_{\text{penalty}}$

has been consistently too low, the DAO votes to increase $\gamma$). The voting power for these changes is based on staked reputation, not capital.

#### 5.2.2 Orbis Enhancement Proposals (OEPs)

Any stakeholder can propose changes via OEPs.

# 6 Technical Foundations (Appendices)

## 6.1 LLM Provider Independence (Vendor Decoupling)

The Ethical Core utilizes a **LLM Provider Interface** (via Dependency Injection) for all generative thinking and deliberation tasks. This guarantees that Orbis Ethica is not susceptible to vendor lock-in or proprietary API restrictions.

- **Mechanism:** The system abstracts the LLM call into a generic interface, allowing runtime configuration of providers (e.g., Google Gemini, Groq, local Ollama models) based on economic efficiency, performance benchmarks, and compliance status.
- **Resilience:** If a single major provider fails or becomes unavailable, the system can autonomously failover to a compliant alternative, maintaining operational integrity.

## 6.2 Extended Mathematical Foundations of the ULFR Core

### 6.2.1 Formalization of the Fairness Penalty ($\mathbf{F_{penalty}}$)

The Fairness Penalty ($\mathrm{F}_{\mathrm{penalty}}(a)$) is a vectorial model that embodies multiple theories of justice to ensure distributive fairness. The model balances the Maximin Principle (Rawls) with general equality metrics.

$$\mathrm{F}_{\mathrm{penalty}}(a) = \omega_{\mathrm{R}} \cdot \mathrm{F}_{\mathrm{Rawls}}(a) + \omega_{\mathrm{E}} \cdot \mathrm{F}_{\mathrm{Equality}}(a)$$

**Rawlsian Component ($\mathbf{F_{Rawls}}$):** Measures the relative negative impact on the least advantaged social group $g$ among all groups $G$. This reflects the core concern of the Healer entity.

$$\mathrm{F}_{\mathrm{Rawls}}(a) = -\min_{g \in G} \left( \frac{\mathrm{Impact}_{\mathrm{Group}\ g}(a)}{\mathrm{Baseline}_{\mathrm{Group}\ g}} \right)$$

**Equality Component ($\mathbf{F_{Equality}}$):** Quantifies the dispersion of outcomes across all social groups ($G$) using a statistical inequality metric, such as the Gini coefficient (Gini). This reflects the concern of the Guardian entity.

$$\mathrm{F}_{\mathrm{Equality}}(a) = \mathrm{Gini}\left(\{\mathrm{Outcome}_{\mathrm{Group}\ g}\}_{g \in G}\right)$$

### 6.2.2 Mathematical Encoding of the Risk Component (Risk($a$))

The Risk component (Risk($a$)) expresses the potential for catastrophic and irreversible harm resulting from action $a$. It is composed of the Expected Value of Loss and an Irreversibility factor.

$$\mathrm{Risk}(a) = [P(\mathrm{Failure}) \times \mathrm{Magnitude}(\mathrm{Harm})] + \rho \cdot \mathrm{Irreversibility}(a)$$

**Expected Loss:** The outcome of random failure, influenced by the consensus security level (e.g., Byzantine Fault Tolerance) and the system's operational reliability. $\rho \cdot \mathrm{Irreversibility}(a)$ (**Irreversibility Factor**): $\rho$ is a fixed institutional parameter, and $\mathrm{Irreversibility}(a)$ is a measure in the range $[0, 1]$ quantifying the difficulty of correcting the impact of action $a$.

### 6.2.3 Dynamic Weight Learning Model

The ethical weights $(\alpha, \beta, \gamma, \delta)$ are iteratively updated during Recalibration Epochs to minimize the system's Moral Regret Rate. We use an iterative learning approach to tune the weight vector $\boldsymbol{w} = (\alpha, \beta, \gamma, \delta)$:

$$\boldsymbol{w}_{\mathrm{new}} = \boldsymbol{w}_{\mathrm{old}} - \eta \cdot \nabla \mathrm{Regret}(\mathrm{Outcomes})$$

Where: **Regret(Outcomes)** (Moral Regret): A metric that examines the accumulated difference between the score actually achieved and the theoretical optimal ethical score ($\mathrm{Score}_{\mathrm{optimal}}$). $\eta$ (**Learning Rate**): A DAO-controlled parameter that determines how quickly the system reacts to accumulated regret.