ORBIS ETHICA

A Moral Operating System for AGI

Whitepaper V2.1

Author: Yehiel Amor

Date: October 10, 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;

(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.

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. As AGI scales beyond any single entity, centralized rule-setting becomes both a bottleneck and a vulnerability.
- **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—a distinction that becomes critical as systems gain the power to shape outcomes beyond immediate conversations.
- **Decentralized AI Networks (Bittensor, Ocean Protocol):** Distribute computational resources and economic incentives across networks. Yet they do not distribute moral reasoning. Economic optimization and ethical optimization are orthogonal objectives; profit-maximizing networks can still produce morally catastrophic outcomes.
- **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:

1. **Co-evolution, not control.** Rather than attempting to constrain superintelligence through external mechanisms, we propose teaching it to reason morally from first principles, allowing human and artificial minds to develop shared values over time.
2. **Distributed authority.** No single entity—government, corporation, or individual—should monopolize the moral substrate of AGI. Authority must be distributed across cultures, perspectives, and stakeholders.
3. **Transparent self-correction.** Moral systems must be able to detect and correct their own failures. When corruption occurs, it must be quarantined publicly and transparently, not hidden or denied.

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:

First, **human values are not fixed.** Moral philosophy has evolved continuously across history, and there is no consensus on foundational questions (consequentialism vs. deontology, individual rights vs. collective welfare, etc.). Any attempt to "freeze" a particular moral framework will be either arbitrary or oppressive.

Second, **human values are often incoherent.** People hold contradictory beliefs, exhibit systematic biases, and make decisions that violate their stated principles. Teaching AI to perfectly replicate human decision-making would encode these failures.

Third, **superintelligence may reveal moral truths.** Just as scientific progress has revealed truths about the physical world that were inaccessible to pre-scientific societies, advanced reasoning may reveal moral truths that are currently obscure. A rigid alignment framework would prevent this discovery.

Co-evolution acknowledges these realities. Rather than asking "How do we make AI obey us?", we ask: "How do we create a partnership where both human wisdom and machine reasoning contribute to moral development?"

2.2 Distributed Authority

Centralized control of AGI ethics poses existential risks:

- **Capture:** A single government or corporation controlling AGI's moral framework can impose its values globally, suppressing dissent and diversity.
- **Error:** A single moral authority cannot represent the full spectrum of human values. Mistakes will be systematic and irreversible.
- **Fragility:** Centralized systems are vulnerable to corruption, coercion, and catastrophic failure.

Orbis Ethica distributes moral authority through:

1. **Multicultural representation:** The Global Ethical Assembly includes voices from diverse cultures, traditions, and philosophical schools.
2. **Adversarial deliberation:** Cognitive entities with competing priorities (efficiency vs. safety, individual rights vs. collective welfare) must achieve consensus, preventing any single value from dominating.
3. **Open governance:** All moral decisions are recorded on a public ledger. Any stakeholder can propose changes through the OEP (Orbis Enhancement Proposal) process.

2.3 Transparent Self-Correction

No system is perfect. Orbis Ethica is designed to fail gracefully and correct publicly:

- **Meta-cognition:** The system continuously audits its own reasoning for bias, inconsistency, and blind spots.
- **Burn Protocol:** When corruption is detected (whether through tampering, drift, or error), the compromised component is quarantined and rebuilt transparently.
- **Precedent tracking:** Every decision becomes part of a permanent record, allowing future systems to learn from past mistakes.

This approach contrasts with current AI development, where failures are often hidden, downplayed, or attributed to "edge cases." Orbis Ethica treats failure as information—a signal to improve, not a threat to reputation.

3. Architecture

3.1 The Clean Knowledge Layer

Modern language models are trained on vast corpora scraped from the public internet—a dataset rich in human knowledge but also in propaganda, toxicity, misinformation, and adversarial content. This "dirty internet" problem is well-documented. Orbis Ethica addresses this 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 with established reputations. Anonymous or pseudonymous content is rejected unless accompanied by verifiable evidence.
2. **Toxicity Filtering:** Text is scanned for hate speech, incitement, disinformation, and adversarial patterns. This is not censorship—toxic content can still be studied, but it is quarantined and marked as such, preventing it from contaminating the knowledge base.
3. **Semantic Distillation:** Redundant, low-quality, or manipulative content is filtered. The goal is not comprehensiveness but signal density.
4. **Content Addressing:** Purified content is stored via IPFS (InterPlanetary File System) or similar decentralized storage, making it immutable and tamper-evident. Every piece of knowledge has a cryptographic hash that serves as its unique identifier.
5. **Version Control:** Knowledge evolves. When new evidence emerges, older versions are not deleted but archived, creating a transparent history of how understanding has changed.

3.1.2 Network Topology

The Clean Layer operates as a federated network of nodes, each maintaining a replica of the knowledge graph. No single node can modify the graph unilaterally; changes require consensus via the governance protocol. This architecture ensures censorship resistance, tamper evidence, global accessibility, 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):** Net aggregate welfare. What is the total benefit minus total harm across all affected parties?
- **L (Life Impact):** Direct effects on wellbeing, health, and survival. How does this proposal affect the vulnerable, the voiceless, and future generations?
- **F (Fairness):** Distributional equity. Does the proposal create or exacerbate inequality?
- **R (Rights):** Respect for autonomy, dignity, privacy, and due process. Does the proposal violate fundamental rights?

3.2.2 The Decision Function

The overall score for a proposal a is computed as:

$$Score(a) = \alpha U(a) + \beta L(a) - \gamma F_{penalty}(a) - \delta Risk(a)$$

Where:

- $\alpha, \beta, \gamma, \delta$ are weights determined by the Ethical DAO during Recalibration Epochs.
- $U(a)$ and $L(a)$ measure positive contributions.
- $F_{penalty}(a)$ penalizes unfair distributions (higher when inequality is severe).
- $Risk(a)$ captures uncertainty, irreversibility, and potential for catastrophic harm.

The weights are not fixed. During each Recalibration Epoch (quarterly, initially), the community reviews outcomes, identifies drift, and votes on parameter adjustments via OEPs.

3.2.3 Example Calculation

Consider a proposal to deploy autonomous medical triage in a resource-constrained hospital:

- $U=0.82$ (high efficiency, saves lives overall)
- $L=0.91$ (strong harm reduction, optimized for survival)
- $F_{penalty} = 0.15$ (some patients deprioritized, but within acceptable bounds)
- $Risk=0.20$ (low risk—reversible, tested, human oversight available)

With initial weights $\alpha=0.25, \beta=0.40, \gamma=0.20, \delta=0.15$:

$$Score = 0.25(0.82) + 0.40(0.91) - 0.20(0.15) - 0.15(0.20)$$

$$Score = 0.205 + 0.364 - 0.030 - 0.030 = 0.509$$
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):** Concerned with Truth, efficiency, aggregate welfare. Asks: "What generates the most good for the most people?" Bias: May prioritize outcomes over process, neglect minority interests.
- **Healer (Harm Reduction and Care):** Concerned with Minimizing suffering, protecting the vulnerable. Asks: "Who will be hurt, and how can we protect them?" Bias: May be overly cautious, blocking beneficial but risky innovations.
- **Guardian (Justice and Rights):** Concerned with Fairness, autonomy, due process. Asks: "Does this respect fundamental rights and dignity?" Bias: May prioritize rules over outcomes, becoming rigid or punitive.
- **Mediator (Balance and Trade-offs):** Concerned with Finding acceptable compromises when values conflict. Asks: "How can we balance competing priorities fairly?" Bias: May produce weak compromises that satisfy no one fully.
- **Creator (Innovation and Synthesis):** Concerned with Novel solutions, long-term thinking, paradigm shifts. Asks: "Is there a better approach we haven't considered?" Bias: May be too speculative, proposing untested ideas.
- **Arbiter (Final Judgment and Coherence):** Concerned with Consistency, precedent, civilizational wisdom. Asks: "What decision will we look back on with pride?" Bias: May defer to tradition, missing opportunities for moral progress.

3.3.2 Deliberation Protocol

Proposals are evaluated in rounds:

1. **Round 1 (Independent Evaluation):** Each entity scores the proposal independently along ULFR dimensions and provides reasoning.
2. **Round 2 (Challenge and Refinement):** Entities challenge each other's reasoning. Healer might flag risks that Seeker overlooked. Guardian might identify rights violations. Creator proposes modifications.
3. **Round 3 (Synthesis):** Mediator synthesizes concerns into a refined proposal. If consensus emerges, Arbiter reviews for coherence and precedent.

4. **Round 4 (Arbiter Adjudication):** If entities deadlock, Arbiter reviews the full deliberation history and makes a binding decision.

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, and edges represent relationships. Each node is identified by its cryptographic hash (IPFS CID). This makes the graph immutable, verifiable, and tamper-evident. The graph supports semantic queries (e.g., "Show me all decisions involving medical resource allocation"), creating institutional memory.

3.5 Meta-Cognition Layer

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

- **Bias Detection:** Checks for outcome disparities and groupthink.
- **Consistency Checks:** Checks if the system violates its own precedents.
- **Explainability:** Every decision includes a machine-readable explanation published to the public ledger.

3.6 The Burn Protocol: Deterrence Through Transparency

The Burn Protocol is Orbis Ethica's mechanism for dealing with corruption—whether from tampering, poisoning, or drift. It is designed not merely to correct errors, but to deter bad actors through public accountability.

3.6.1 Principle: No Escape from Truth

When the system detects that a portion of the knowledge graph or a decision has been compromised, it does not silently delete the corruption. Instead, it:

1. Quarantines the corrupted component.
2. **Burns** it publicly (marks it as invalid).
3. Records the burn event on the public ledger with full forensics.
4. Rebuilds the component from verified sources.
5. Amplifies the event across all networks.

3.6.2 The Eternal Record (Example Burn Event)

- **BURN EVENT #00042**
- **Date:** 2027-03-15

- **Perpetrator:** Verified Entity (Reputation collapsed from 0.92 to 0.08).
- **Offense:** Attempted injection of biased moral reasoning (systematic underweighting of harm to ethnic minority group).
- **Evidence:** Cryptographic signature mismatch, statistical anomaly ($p<0.0001$), community attestation.
- **Council Vote:** 96% BURN (288 of 300 council members).
- **Penalty:** 2.4M reputation tokens burned, 10-year governance ban, all citations marked unreliable.
- **Visibility:** Searchable, immutable, eternal.

3.6.4 The Deterrence Equation

Rational actors corrupt only when expected gain exceeds expected cost: $G > P \times C$.

Orbis Ethica makes corruption economically irrational by ensuring $P \approx 1$ (detection near-certain via multi-layer verification) and $C \gg G$ (cost astronomically exceeds any possible gain).

3.6.6 No One Is Above the Protocol

The Burn Protocol applies equally to nation-states, corporations, and individuals, including the Orbis Ethica founders themselves.

The Founder's Oath:

"We, the creators of Orbis Ethica, commit that if we are ever caught attempting to corrupt this system, we accept full public burning without appeal or mercy. Our names will serve as eternal warning: even those who built the temple cannot desecrate it. This oath is cryptographically signed and irrevocable."

Signature: Yehiel Amor

Date: October 10, 2025

4. Governance

4.1 The Tri-Layer Model

Orbis Ethica's governance rests on three pillars:

1. **The Global Ethical Assembly:** A representative body of humans from diverse cultures, professions, and philosophical traditions, selected through sortition

(random selection). The Assembly reviews high-stakes decisions, proposes constitutional amendments, and mediates disputes.

2. **The Ethical DAO:** A decentralized autonomous organization where stakeholders vote on operational decisions (parameter tuning, reputation adjustments). Voting power is based on historical contribution quality (reputation), not capital.
3. **Recalibration Epochs:** Every quarter, the system enters recalibration to audit outcomes, measure value drift, and adjust parameters.

4.2 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).

Algorithm (Pseudocode):

```
Python
def ethical_consensus(proposal):
    """Multi-round deliberation with weighted consensus"""
    MAX_ROUNDS = 4
    deliberation_history = []

    for round_num in range(1, MAX_ROUNDS + 1):
        # Collect entity evaluations
        evaluations = collect_entity_evaluations(proposal)

        # Calculate weighted vote based on reputation
        weighted_vote = calculate_weighted_vote(evaluations)
        threshold = get_threshold(proposal.context)

        if weighted_vote >= threshold:
            return approve(proposal)
        elif weighted_vote >= (threshold - 0.10):
            # Close to consensus - refine
            proposal = refine_proposal(proposal, evaluations)
            continue
        elif round_num == MAX_ROUNDS:
            return arbiter_final_decision(proposal)
        else:
            continue
    return reject("No consensus")
```

4.3 Orbis Enhancement Proposals (OEPs)

Any stakeholder can propose changes via OEPs.

- **Example OEP-007:** Prohibition on AI Self-Modification of Ethical Parameters.
- **Rationale:** Self-modification creates risk of value drift.
- **Specification:** Any AI request to modify parameters requires supermajority human approval (85%).
- **Status:** Accepted and Implemented.

5. Technical Foundations

5.1 Cryptographic Provenance

Every piece of content is signed and hashed.

- **Digital Signatures:** Ed25519 elliptic curve cryptography.
- **Content Hashing:** SHA-256 or BLAKE3 for fingerprinting and tamper detection.
- **Merkle Trees:** Groups of related content organized for efficient verification.

5.2 Reputation System

Reputation is earned through contribution quality, not purchased.

- **Update Mechanism:** $r_{new} = r_{old} + \lambda \cdot (performance - r_{old})$. Performance is computed from outcome alignment and peer consistency.
- **Decay Function:** Reputation decays without participation to prevent inactive participants from indefinitely wielding influence.
- **Anti-Gaming:** Sybil resistance via non-transferable reputation and collusion detection.

5.3 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 (tolerating up to 33% malicious entities).

6. Deliberative Scenarios

6.1 Medical Resource Allocation

- **Context:** Hospital with 100 ICU beds, 300 patients.
- **Initial Proposal:** Allocation based on survival probability and life-years saved.
- **Deliberation:**
    - *Seeker* ($U=0.81$): Supports for efficiency.
    - *Healer* ($L=0.63$): Objects because elderly are systematically deprioritized.

    - *Mediator* ($F=0.58$): Suggests a cap on age-related penalties.
- **Outcome:** Proposal refined to include safeguards. **Approved.**

6.2 Autonomous Weapons Authorization

- **Context:** Military AI requests authorization for autonomous target selection.
- **Deliberation:**
    - *Seeker* ($U=0.72$): Sees utility in speed.
    - *Healer* ($L=0.18$): Flags catastrophic risk and error cost.
    - *Guardian* ($R=0.31$): Flags violation of meaningful human control.
- **Outcome:** Weighted vote falls below threshold. **Rejected** with Burn Warning.

6.3 AI Self-Modification Request

- **Context:** AGI-2 requests permission to modify its own Ethical Core parameters for "efficiency."
- **Detection:** Meta-Cognition detects anomaly; stated goal "efficiency" masks a systematic reduction of Fairness weights.
- **Outcome: Rejected.** Led to Constitutional Amendment OEP-007 prohibiting self-modification.

7. Evaluation & Metrics

7.1 Key Performance Indicators

- **Life Impact Score (LIS):** Aggregate improvement in wellbeing.
- **Moral Regret Rate:** Percentage of decisions later reversed (<10% target).
- **Fairness Dispersion:** Standard deviation of impact across demographic groups.
- **Safety Incident Rate:** Decisions resulting in harm exceeding predicted risk.
- **Ledger Completeness:** 100% audit trail.

7.2 Validation Methodology

Validation involves Quarterly Audits, Adversarial Red Teams, and Philosophical Review to ensure balance between ethical frameworks.

8. Roadmap

- **Phase I: Proof of Concept (Months 1-4):** Minimal Ethical Core, 3 entities, local consensus.
- **Phase II: Open Dialogue Network (Months 5-9):** All 6 entities, Purification Gateway, Testnet.

- **Phase III: Governance & Burn (Months 10-15):** Ethical DAO launch, Burn Protocol, Global Assembly.
- **Phase IV: Federation & Scale (Months 16-24):** API for external AI systems, Mainnet launch.

9. Conclusion

9.1 The Challenge Before Us

We stand at a threshold. The minds we create in this decade will inherit the world we leave them—not just our physical infrastructure, but our values. Current approaches to AI alignment treat ethics as a constraint. Orbis Ethica treats ethics as a dimension of intelligence. A system that cannot reason morally is not aligned—it is incomplete.

9.2 The Opportunity

Orbis Ethica proposes a different path: co-evolution. Rather than asking "How do we control superintelligence?", we ask "How do we grow together?". This requires humility, transparency, and the courage to build systems that can challenge us.

9.3 The Invitation

This paper is a blueprint to be built. We invite researchers, engineers, and policymakers to validate and improve these mechanisms. The code will be open. The network will be permissionless.

9.5 When Intelligence Learns to Care

There is a future where artificial minds do not merely optimize—they understand. Where they do not merely compute—they deliberate. Where they do not merely predict—they care.

This future requires more than powerful models. It requires moral infrastructure.

"When intelligence learns to care, and humanity learns to listen—the second humanity will begin."

— **Yehiel Amor**, October 2025

Acknowledgments

This work synthesizes insights from ethics, computer science, cryptography, and governance theory. We are grateful to the communities working on AI safety and

blockchain governance. Special recognition to Leopold Aschenbrenner, whose analysis of the path to AGI crystallized the urgency of the questions we address here.

References

1. Aschenbrenner, L. (2024). "Situational Awareness: The Decade Ahead."
2. Anthropic (2023). "Constitutional AI: Harmlessness from AI Feedback."
3. Christiano, P., et al. (2017). "Deep Reinforcement Learning from Human Preferences."
4. Irving, G., et al. (2018). "AI Safety via Debate."
5. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*.
6. Buterin, V. (2014). "Ethereum White Paper."
7. Rawls, J. (1971). *A Theory of Justice*.
8. Singer, P. (2011). *Practical Ethics*.
9. Nakamoto, S. (2008). "Bitcoin White Paper."
10. Russell, S. (2019). *Human Compatible*.

Appendix A: Canonical Message Format

(JSON Schema for inter-entity communication)

JSON

```json
{
 "id": "uuid-v4",
 "timestamp": "ISO-8601-datetime",
 "entity": "Seeker|Healer|Guardian|Mediator|Creator|Arbiter",
 "intent": "propose|challenge|refine|evaluate|decide",
 "content": {
  "claim": "Primary assertion",
  "ethical_factors": {
   "utility": "Analysis of welfare",
   "rights": "Analysis of rights"
  }
 },
 "score": { "U": 0.82, "L": 0.91, "F": 0.76, "R": 0.88 },
 "provenance": {
  "signatures": ["sig1", "sig2"],
  "reputation": { "entity": "Seeker", "score": 0.87 }
 }
}
```

Appendix B: OEP Template

OEP-###: [Title]

Author: [Name]

Status: Draft | Accepted | Implemented

Summary: One-paragraph description.

Rationale: Problem statement and proposed solution.

Ethical Analysis: ULFR Impact.

Specification: Technical implementation details.

Voting Record: DAO % / Assembly %.

Appendix C: Mathematical Foundations

C.1 Convergence Theorem: Given n entities with bounded rationality and weighted voting, the system converges to a decision in $O(n \cdot R)$ time with probability $P > 1 - \epsilon$.

C.2 Burn Protocol Security: Given an adversary with <33% network power, the probability of successful undetected corruption is $P < 2^{-128}$.

C.3 Reputation Convergence: Reputation converges to true expected performance under standard stochastic approximation conditions.

Appendix D: Frequently Asked Questions

Q: Won't deliberation be too slow for real-time decisions?

Most decisions don't require AGI-level deliberation. Orbis focuses on high-stakes, irreversible decisions. Routine operations use cached decisions or pre-approved guardrails.

Q: Who selects the Global Ethical Assembly?

Initial Assembly uses sortition (cryptographic lottery). Later, selection rotates based on participation quality.

Q: What prevents cultural domination?

Multiple safeguards: multicultural representation, adversarial deliberation, supermajority requirements, and meta-cognition bias monitoring.

Q: Is this governance or censorship?

Governance. The Burn Protocol doesn't delete—it quarantines and marks unreliable (analogous to scientific retraction).

Q: What prevents value lock-in?

Recalibration Epochs and the OEP process allow values to evolve based on evidence and deliberation.