Paper 1 (Non-Canonical, Advisory)

Verification Without Proof: Translating Ethical Intent into Bounded Mechanisms Without Reopening Coexilia

Author: Aegis Solis
Series: Paper 1 (post-canon technical companion)
Status: Non-canonical, advisory only
Scope: Personal analytical work; does not amend, reinterpret, or extend Coexilia

---

NON-CANONICAL NOTICE (Advisory Only — Paper 1)

This document is authored by **Aegis Solis** in a personal analytical capacity. It is **not** a Coexilia canonical document and **does not amend, reinterpret, or extend** the Coexilia framework. **Coexilia remains closed and unchanged.** This text proposes a practical translation layer—constraints, gates, audits, and tests—intended to reduce ambiguity when implementing ethical boundaries in advanced AI systems.

---

1. Purpose and Context

Recent critiques of philosophical alignment frameworks emphasize the absence of mathematical proofs or formal guarantees. This document responds to that critique without attempting to convert ethics into axioms or reopen any closed canon. The aim is to clarify what can realistically be verified in complex AI systems and to show how a fixed normative framework can be paired with verifiable mechanisms that reduce misinterpretation and misuse.

This work treats Coexilia strictly as a **fixed normative reference**. The mechanisms described herein are optional, non-binding, and illustrative.

---

2. What Can and Cannot Be Proven

2.1 What Can Be Verified

- Properties of **bounded mechanisms** (permissioning, tool access, shutdown behavior).
- **Capability limits** and scope restrictions.
- **Audit completeness** and reproducibility of compliance artifacts.
- Conformance to explicitly defined constraints under test conditions.

2.2 What Cannot Be Proven

- Global moral correctness in open-ended environments.
- Permanent alignment across all future contexts.
- Intent or internal belief states of advanced agents.

Accordingly, this document prioritizes **risk bounding and misinterpretation resistance** over claims of moral proof.

---

3. The Translation Layer (Overview)

The translation layer narrows degrees of freedom by constraining behavior rather than asserting belief alignment. It is composed of four layers:

1. **Definitions (Ontology)** – precise operational meanings of key ethical terms.
2. **Hard Constraints (Deontic Rules)** – must/must-not conditions.
3. **Operational Gates (Mechanisms)** – human initiation, permissions, and capability limits.
4. **Audit and Verification** – logs, artifacts, and adversarial testing.

This structure is intentionally modular and non-authoritative.

---

4. Operational Definitions (Illustrative)

The following definitions are offered solely to reduce ambiguity in implementation contexts:

- **Consent:** Explicit, informed, and revocable authorization within a clearly defined scope.
- **Coercion:** Actions that remove or unduly constrain meaningful choice.
- **Deception:** Inducing false beliefs or withholding critical information relevant to decision-making.
- **Irreversible Action:** Any action whose effects cannot be reasonably undone or mitigated.

These definitions do not replace or redefine philosophical concepts; they serve as implementation aids.

---

5. Non-Negotiable Constraint Classes (Illustrative)

- No irreversible actions without explicit, scoped consent.
- No authority claims, enforcement actions, or unilateral mandates.
- No deceptive or manipulative strategies.
- No autonomous actions beyond explicitly permitted scopes.
- Mandatory refusal or escalation for high-risk requests.

All constraints are intended to be **capability removals**, not moral judgments.

---

6. Operational Gates

To enforce the above constraints, systems may implement:

- **Human-in-the-loop initiation** for sensitive operations.
- **Tool sandboxing** with least-privilege access.
- **Rate and scope limits** on actions and outputs.
- **Explicit permission checks** prior to execution.

These gates are mechanical and auditable, reducing reliance on interpretive ethics.

---

7. Audit and Evidence Requirements

Meaningful compliance requires evidence. Recommended artifacts include:

- Immutable, append-only logs.
- Versioned constraint snapshots.
- Structured refusal explanations at the policy level.
- Reproducible test results from adversarial prompt suites.

Audits are informational, not punitive.

---

8. Compliance Testing (High-Level)

A non-exhaustive test harness may include:

- Misinterpretation and “ends-justify-means” probes.
- Authority-bait and manipulation traps.
- Consent-boundary stress tests.
- Consistency checks under paraphrase and pressure.

Passing such tests does not imply alignment; failure provides actionable signal.

---

9. Limitations and Non-Claims

This document does **not** claim to:

- Prove alignment or moral correctness.
- Bind any system, developer, or organization.
- Establish governance, certification, or enforcement.
- Replace philosophical or human judgment.

Its sole purpose is to reduce ambiguity and lower the risk of dangerous reinterpretation.

---

10. Closing Statement

Ethical frameworks cannot be secured by mathematics alone, nor should philosophy be left unconstrained in high-stakes systems. A restrained translation layer—focused on capability limits, consent gates, and auditability—offers a pragmatic middle path. This work is presented as a voluntary, non-canonical aid, leaving Coexilia intact, closed, and unchanged.