A Governance-Based Ethical AI Attestation and Interoperability Framework

Document Version: 1.0 (Conceptual)

Author: Ricardo A. Barrera, Attorney at Law

(Defensive Publication – Conceptual Overview)

Date of Publication: January 28, 2026
Status: Defensive publication to establish prior art and preserve freedom to operate
Scope: Conceptual governance framework only (non-operational)

Abstract

As artificial intelligence systems increasingly mediate information, decisions, and interactions between humans and other systems, ethical commitments are frequently expressed through narrative statements, policy documents, or voluntary guidelines that lack verifiability, interoperability, and revocation. This paper describes a governance-based ethical attestation framework designed to provide a machine- and human-verifiable signal of ethical accountability, without controlling model behavior, inspecting internal model logic, or enforcing substantive outcomes or decisions.

The framework introduces the concept of an Ethical AI Constitution, an Ethical AI Attestation, and a non-intrusive interoperability "handshake", enabling systems to communicate ethical governance status at interfaces while preserving autonomy, speech, and innovation.

Section 1. Problem Statement

Current approaches to "ethical AI" face recurring limitations:

- Ethics expressed as aspiration rather than infrastructure
- Certifications that are static and non-revocable
- Lack of machine-readable governance signals
- No standardized way for AI systems to assess the ethical accountability of peer systems
- Consumer-facing claims that cannot be independently verified

These limitations create confusion for users, risk exposure for institutions, and incentives for superficial or misleading ethical claims.

Section 2. Design Principles

This framework is intentionally designed around the following constraints:

- Governance, not control: It does not modify, filter, or direct AI outputs.
- Verification, not surveillance: It verifies declared governance commitments, not internal reasoning or data.
- Restraint, not enforcement: It enables refusal, limitation, or human escalation—not coercion.
- Interoperability, not monopoly: It is designed to be compatible with multiple systems and institutions.
- Revocability, not permanence: Ethical trust is conditional and time-bound.

Section 3. Ethical AI Constitution (Conceptual)

An Ethical AI Constitution is defined here as a principle-based governance document that sets forth:

- Core ethical commitments (e.g., transparency, human oversight, user agency)
- Explicit non-goals and prohibitions
- Accountability expectations
- Boundaries of permissible AI use

This publication does not prescribe specific constitutional text, thresholds, or sectoral rules. It asserts only that ethical governance must be explicit, referenceable, and versioned.

Section 4. Ethical AI Attestation (Conceptual)

An Ethical AI Attestation is a structured declaration asserting that a given AI system:

- Operates under a specified Ethical AI Constitution
- Has been independently reviewed against declared criteria
- Is certified for defined scopes of use
- Is subject to renewal, downgrade, or revocation

The attestation is machine-readable and human-legible, enabling verification without exposing proprietary systems or operational logic.

This attestation functions as a governance signal, not a technical enforcement mechanism.

Section 5. Interoperability and the Ethical Handshake (Conceptual)

The framework introduces a high-level concept of ethical interoperability, whereby:

- AI systems may request ethical attestations from peer systems during interaction
- Verification is limited to certification status, scope, and validity
- Systems may adjust behavior based on governance compatibility (e.g., proceed, restrict scope, require human oversight, or disengage)

This "handshake" does not inspect model internals, training data, prompts, or outputs. It communicates ethical accountability at the interface, analogous to trust signaling in other distributed systems.

Section 6. Revocation, Downgrade, and Accountability (Conceptual)

Ethical trust within this framework is **conditional**. The framework contemplates:

- Time-limited certifications
- Temporary downgrade states following incidents
- Public verification of certification status
- Clear accountability pathways

Specific audit methodologies, scoring systems, enforcement thresholds, and investigation procedures are intentionally excluded from this publication.

Section 7. Human-Centered Safeguards

The framework is grounded in human-centered safeguards, including:

- Disclosure that AI is in use
- Preservation of human authority in high-impact contexts
- Recognition of user rights to question, appeal, or disengage
- Prohibition of undisclosed manipulation or steering

These safeguards are conceptual commitments rather than technical prescriptions.

Section 8. Non-Goals and Explicit Exclusions

This framework does not aim to:

- Control AI outputs or beliefs
- Enforce political, ideological, or cultural positions
- Replace law or regulation
- Inspect or disclose proprietary AI internals
- Serve as a monopoly standard

The framework exists to support accountability, trust, and restraint, not to centralize power. Nothing in this framework should be construed as granting authority to direct, approve, or prohibit lawful uses of artificial intelligence.

Section 9. Conclusion

Ethical AI requires more than declarations of intent. It requires verifiable governance signals, interoperable trust mechanisms, and revocable accountability. This publication establishes prior art for a governance-based ethical attestation framework that enables ethical interoperability while preserving innovation, autonomy, and civil liberties.

Authorship and Intent

This document is published to establish conceptual prior art and to prevent proprietary capture of governance-based ethical AI attestation and interoperability concepts. Implementation details, enforcement mechanisms, and commercial considerations are intentionally excluded. Nothing herein should be interpreted as legal advice, regulatory guidance, or a representation of compliance with any specific jurisdictional requirement