draft-reilly-uaemf-00

Internet Engineering Task Force (IETF)
Internet-Draft
Intended status: Informational
Expires: September 2026

L. Reilly
Independent

March 2026

Universal AI Ethics and Moral Framework (UAEMF)
The Moral Compass of Artificial Intelligence

Abstract

This document provides a detailed explanatory rendering of the Universal AI Ethics and Moral Framework, abbreviated UAEMF. The framework is presented as a universal moral architecture for the governance of artificial intelligence systems. It is designed not merely as a short ethics statement, but as a structured document that moves from first principles to practical obligations.

The explanatory goal of this draft is to describe what the framework is, what problem it is trying to solve, how its axioms function, how its twelve principles are meant to operate, why its tiered governance structure matters, and how its permanence and provenance mechanisms support integrity and historical traceability.

The source framework is archived through Zenodo under DOI 10.5281/zenodo.19010455 and is represented as cryptographically timestamped through OpenTimestamps in Bitcoin block 940570, with the attestation date shown as 2026-03-13 EST.

1. Introduction

Artificial intelligence systems now shape outcomes that were once controlled only by human decision makers. AI influences who is hired, what information is ranked, how credit is assessed, how content is generated, how safety signals are interpreted, and in some cases how state or institutional power is exercised. Because AI can affect rights, opportunities, reputation, safety, and autonomy at scale, a framework for ethical governance is needed.

The Universal AI Ethics and Moral Framework, or UAEMF, is intended to meet that need by acting as a moral compass. The central claim of the framework is that AI governance should not begin and end with technical performance or market utility. Instead, it should be anchored in human dignity, accountability, meaningful consent, transparency, non-discrimination, safety, and long-term human flourishing.

UAEMF is therefore best understood as a constitutional style framework for AI ethics. It begins with foundational axioms, expands into universal principles, translates those principles into practical duties, identifies absolute prohibitions, and proposes governance and compliance structures for systems of different risk levels.

2. What the UAEMF Document Is

The UAEMF document is a comprehensive ethical governance framework for artificial intelligence. It is broader than a vendor policy, broader than a single law, and broader than a technical standard that focuses only on implementation details. Its purpose is to establish a moral reference point that can guide developers, organizations, regulators, and institutions regardless of country or sector.

The framework contains several layers. First, it contains a declaration of purpose that explains why AI requires a moral compass. Second, it contains foundational axioms that define the moral basis of the framework. Third, it defines its scope, including what counts as AI, what counts as a consequential decision, and what kinds of domains count as high stakes. Fourth, it sets out twelve universal principles. Fifth, it translates those principles into stakeholder obligations, implementation standards, and red-line prohibitions. Sixth, it describes domain-specific applications and a multi-tier compliance architecture. Finally, it reinforces the integrity of the document through archival and timestamping mechanisms.

This means the document is both philosophical and operational. It contains claims about what human beings are owed in the age of AI, and it also contains claims about how institutions should behave if they wish to act ethically in that age.

3. Foundational Ethical Axioms

UAEMF begins with three foundational axioms. These axioms are not peripheral. They are the deepest layer of the framework. Each later principle is meant to flow from them.

3.1. The Dignity Axiom

The Dignity Axiom states that human beings possess inherent worth that is not dependent on productivity, utility, compliance, or status. In AI governance, this means a system must never treat people merely as computational objects to be scored, sorted, manipulated, or optimized. Dignity is not satisfied merely by avoiding physical harm. The concept also includes psychological integrity, freedom from dehumanization, and respect for the person as a full moral subject.

3.2. The Accountability Axiom

The Accountability Axiom states that power exercised over human beings

without accountability, transparency, and meaningful contestation is illegitimate. In practice, this means responsibility for AI outcomes cannot evaporate into code or model complexity. Someone remains answerable for the decision to build, deploy, approve, or fail to control the system. The axiom is the basis for audit trails, named responsible parties, incident reporting, and remedy paths for harm.

3.3. The Consent Axiom

The Consent Axiom states that legitimate AI use upon or about individuals depends on genuine, informed, specific, freely given, and revocable consent. It rejects manufactured consent, buried consent, or consent coerced by the threat of losing essential services. The axiom is especially important in contexts involving data collection, profiling, training data use, and automated decisions with personal impact.

Together, the three axioms establish a simple but powerful thesis. AI may be advanced, efficient, and profitable, but it is not ethically acceptable unless it preserves human dignity, remains accountable to human institutions, and respects meaningful human consent.

4. Scope and Reach of the Framework

UAEMF defines artificial intelligence broadly enough to include machine learning, statistical inference, large language models, computer vision, natural language processing, robotics, and related computational methods when they perform tasks historically associated with human cognition or decision support.

The framework also defines an AI system broadly. It is not limited to a model file or algorithm. It includes the model, the training data, the inference infrastructure, the operational environment, and the deployment context that gives the output real-world force.

Another important concept is the consequential decision. The document treats as consequential any AI-driven output that materially affects access to employment, housing, healthcare, education, credit, legal status, public services, physical safety, or freedom. This definition matters because it means the framework is aimed at the actual human stakes of AI deployment rather than at technical novelty alone.

The framework claims lifecycle scope and global scope. It applies from early design through data gathering, training, testing, deployment, monitoring, and decommissioning. It also claims that weak local regulation does not cancel moral obligations. In that sense, the document is intentionally universal in ambition.

5. The Twelve Universal Principles

The twelve principles are the operational core of UAEMF. They are not presented as a menu of optional values. The framework says they form an integrated moral architecture, meaning they are supposed to work together and reinforce one another.

5.1. Principle 1: Human Dignity and Non-Subjugation

This principle states that no efficiency gain, profit motive, power interest, or national objective supersedes the worth of a human being. It opposes systems that reduce people to profiles, case numbers, risk scores, or manipulable emotional targets. It treats dehumanization as a primary moral failure of AI.

5.2. Principle 2: Transparency and Explainability

This principle states that when AI contributes to a consequential decision, people should be able to understand that AI was used, what

factors influenced the outcome, and how to contest it. The emphasis is on human-comprehensible explanation rather than merely technical description. This principle also supports disclosure of AI-generated or AI-modified content in contexts where truth and authorship matter.

5.3. Principle 3: Accountability and Answerability

This principle states that organizations cannot escape responsibility by saying the model made the decision. Moral and institutional responsibility remains with the humans and institutions that chose to design, approve, deploy, or supervise the system. Accordingly, the principle supports named accountable parties, decision logs, incident disclosure, and remediation paths.

5.4. Principle 4: Privacy and Data Sovereignty

This principle treats personal data as an extension of identity and self-determination. It rejects silent extraction, broad repurposing, and the use of personal information for AI training or profiling without meaningful consent. It is especially strict regarding biometric data, because such data is intimate, identifying, and often irreversible once exposed.

5.5. Principle 5: Equity, Fairness, and Non-Discrimination

This principle says that algorithmic discrimination is still discrimination even when it is expressed through statistics rather than direct animus. Because historical data can encode historical injustice, the framework treats fairness analysis, auditability, intersectional review, and disparity testing as necessary safeguards.

5.6. Principle 6: Safety, Security, and Non-Maleficence

This principle extends the duty to avoid harm into the AI domain. It covers robustness, exploit resistance, adversarial resilience, monitoring, and the ability to pause or withdraw systems when failure or misuse becomes likely. It rejects the pattern of deploying first and apologizing later in high-stakes environments.

5.7. Principle 7: Human Autonomy and the Right to Opt Out

This principle states that the right to refuse certain forms of AI governance must be genuine rather than illusory. If declining algorithmic treatment leads to punishment, exclusion, or denial of essential services, then the apparent consent is not genuinely free. This principle therefore supports meaningful opt-out and meaningful human review.

5.8. Principle 8: Democratic Integrity and Resistance to Capture

This principle moves beyond individual rights and focuses on institutional legitimacy. It warns that AI can distort democratic life through political manipulation, disinformation amplification, epistemic concentration, or capture of regulators by the very industries they are meant to govern. It therefore treats electoral and governance uses of AI as especially sensitive.

5.9. Principle 9: Intellectual Integrity and the Right to Truth

This principle addresses the information environment. It argues that truth is a practical precondition for journalism, science, law, democratic choice, and informed consent. The framework therefore supports provenance, attribution, and strong controls against fabricated evidence, deceptive impersonation, and synthetic fraud.

5.10. Principle 10: Human Oversight and the Prohibition of Unchecked

Autonomy

This principle states that an AI system that cannot be corrected, overridden, paused, or shut down by authorized humans has moved beyond the status of a controllable tool. The framework does not reject automation, but it rejects consequential autonomy without meaningful human supervision and override capacity.

5.11. Principle 11: Children and Vulnerable Population Protection

This principle provides heightened protection for those least able to defend themselves in technologically mediated environments. It covers children, those in crisis, those with impaired judgment, and others exposed to coercive or exploitative conditions. The framework treats manipulation of vulnerability as among the gravest ethical failures.

5.12. Principle 12: Environmental Stewardship and Intergenerational Justice

This principle expands concern to future generations. It argues that AI development should not consume resources, create irreversible risks, or concentrate power in ways that diminish the options, safety, and self-determination of those who come later. In this sense the framework extends from immediate human effects to long-term civilizational responsibility.

6. Governance Architecture and Compliance Logic

UAEMF does not stop at abstract ethical language. It also attempts to organize compliance through a tiered governance structure. The framework description associated with the document identifies a five-tier architecture designed to scale obligations according to the stakes of the system.

At the highest severity end, Tier 0 is reserved for systems or uses that should be prohibited outright, such as autonomous lethal targeting, coercive social scoring, certain forms of mass surveillance, and synthetic child sexual abuse material. Tier 1 covers the highest-risk systems that may be heavily restricted or require stringent safeguards, including criminal justice AI, election systems, child welfare systems, and certain clinical or military applications. Lower tiers correspond to high-impact commercial uses, moderate-risk systems, and comparatively low-risk systems.

This architecture matters because it shows the framework trying to balance universality with proportionality. It does not pretend that every autocomplete tool and every parole recommendation engine should face identical scrutiny. Instead, the framework uses the same moral compass while scaling compliance expectations with the magnitude of possible harm.

7. Provenance, Zenodo, and Blockchain Timestamping

A distinctive feature of UAEMF is the emphasis it places on permanence. The framework argues that many ethics statements fail because they can be quietly revised, softened under pressure, or later presented as if they had always existed in a stronger form. To resist that pattern, the document is tied to both a public archive and a cryptographic timestamp.

The public archival layer is the Zenodo record under DOI 10.5281/zenodo.19010455. Zenodo functions as a stable scholarly repository and provides a persistent identifier for the work. That supports citation, retrieval, and bibliographic continuity.

The second layer is the OpenTimestamps attestation represented as anchored in Bitcoin block 940570, with the attestation date shown as

2026-03-13 EST. The role of this layer is not to prove the truth of the framework’s claims. Its role is to support chronology and integrity by showing that the document existed in a particular form at or before the recorded block height.

Together, the DOI archive and blockchain attestation form a two-layer provenance model: one scholarly and public-facing, the other cryptographic and temporal.

8. IETF Rendering and Submission Considerations

This rendering is formatted as a plaintext Internet-Draft style document. Because Datatracker and idnits checks are sensitive to exact boilerplate wording, line length, first-page naming, and character encoding, this version preserves the draft name on the first line, uses ASCII-only text, wraps content to approximately 72 columns, and includes the standard Internet-Draft and copyright boilerplate language.

ASCII-only rendering is used here intentionally. Although richer characters can be visually appealing, Internet-Draft validation commonly expects plain ASCII for the plaintext submission path. That means the correct practical form for this file is without non-ASCII characters, even when the underlying content is extensive and detailed. This is done to support successful submission and clean validation.

9. Security Considerations

The subject matter of this document is not a wire protocol, but it has direct security relevance. AI systems may be exploited through adversarial prompts, poisoned training data, model extraction, unsafe tool invocation, deceptive content generation, and misuse of apparently harmless features. Ethical governance therefore has a security dimension.

The framework described here supports security not merely through technical hardening but through governance mechanisms such as logging, monitoring, red-teaming, incident reporting, human override, and clear assignment of institutional responsibility.

A further concern is ethics washing. A framework can be cited as evidence of responsibility while being ignored in practice. For that reason, the UAEMF emphasis on traceability, auditability, and public accountability is also a form of governance security.

10. IANA Considerations

This document has no IANA actions.

11. References

11.1. Informative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174.

[RFC5378] Bradner, S. and J. Contreras, "Rights Contributors Provide to the IETF Trust", RFC 5378.

[RFC7322] Flanagan, H. and N. Brownlee, "RFC Style Guide", RFC 7322.

[BCP78] IETF Trust, "Legal Provisions Relating to IETF Documents".

[BCP79] IETF, "Intellectual Property Rights in IETF Technology".

[ZENODO] Reilly, L., "Universal AI Ethics and Moral Framework (UAEMF) v1.0", Zenodo, DOI 10.5281/zenodo.19010455.

Author’s Address

Lawrence Reilly

Independent Researcher

Email: Lawrencejohnreilly@gmail.com