Deriving the 12 Canonical Values of the AI Moral Code: A Stratified Semantic Framework (2006–2025)

Randy J. Hinrichs and Sharon Stoll, PhD.

Abstract

This technical report outlines the empirical derivation of twelve canonical values[1] that serve as the core of the AI Moral Code. Building upon the broader architecture introduced in Hinrichs (2025), *Advancing Ethical AI: A Methodological and Empirical Approach to the AI Moral Code*, this paper provides a focused analysis of value emergence using a stratified semantic framework and composite scoring models. The methodology integrates frequency-weighted scoring, sectoral normalization, and contextual multipliers to isolate and weight the values most frequently and meaningfully cited across 291+ global AI ethics documents spanning from 2006 to 2025. This canonical set functions as the normative seed for value-aligned AI governance and supports continuous recalibration via a proposed Ethical Salience Tracker (EST)[2].

Introduction

The AI Moral Code project seeks to operationalize a universal yet adaptive ethical foundation for artificial intelligence. While the ICAD paper introduced the full NRBC framework—Normative, Regulatory, Behavioral, and Conceptual domains—this

The term canonical refers to values that consistently emerge that consistently emerge across the AI ethics landscape between 2006 and 2025 with high frequency, sectoral breadth, and ethical relevance. In this framework, canonical values are not universal claims, but empirically grounded reference points for human-AI moral partnership, subject to future refinement.

[1] Canonical, as used here, does not imply fixed or universal values, but those with consistent ethical relevance, frequency, and institutional recognition across 291+ AI ethics documents (2006–2025).

[2] The Ethical Salience Tracker (EST) is a temporal recalibration model that quantifies the relative ethical salience of canonical AI values over time. It incorporates document frequency trends, sectoral weight shifts, and emerging thematic proximities to update value rankings dynamically, functioning analogously to a moral volatility index for AI governance.

document focuses exclusively on the derivation of the canonical value set that anchors the Normative layer. Values are selected not on the basis of prescriptive tradition, but through an empirical analysis of value prevalence, conceptual relevance, and sectoral distribution. Value prevalence refers to the measurable frequency and distribution of an ethical value across a defined corpus of documents, sectors, or temporal spans. In the context of AI ethics, it captures how often a particular value (e.g., fairness, accountability) appears—explicitly or through its cognates—within policy texts, principles, or technical guidelines. It serves as a proxy for that value's normative visibility, stakeholder prioritization, and conceptual anchoring.

In this framework, value prevalence is computed through stratified term frequency (TF), adjusted by inverse document frequency (IDF) and sector weighting, forming the empirical foundation for inclusion in the canonical set.

This report assumes familiarity with the corpus construction, NRBC design logic, and simulation framework already published in Hinrichs (2025). Here, the focus narrows to the computational and semantic extraction of a durable, ethically grounded value core.

Methodological Overview

Corpus Construction

291+ AI ethics documents were collected and classified by sector (Government, Industry, Academia, NGO, Religious Organization) and semantically embedded using advanced NLP methods. Sentence-BERT and transformer-based models enabled sentence and paragraph-level vectorization. The curated corpus spans the most influential frameworks, including those by the European Commission (European Commission, 2019), Google (Google, 2018; Google, 2021;

Google, 2023; Pichai, 2018), Floridi et al. (2018), Access Now (2018), and the Pontifical Academy for Life (Pontifical Academy for Life, 2020).

Each document was evaluated not only for linguistic content but for normative clarity, sector-specific expectations, and ethical transparency. This approach embeds our canonical values—such as dignity, fairness, accountability, and trust—into the evaluative logic. For example, documents lacking operational accountability clauses (e.g., audit trails, revision protocols) scored lower in simulation resilience, reflecting the conditional role of accountability.

Key Cross-Sectoral Expectations in AI Ethics & Data Analytics Culture

Across all five sectors, minimum ethical maturity required:

- **Defined Value Sets**: Ethical frameworks must explicitly name guiding values (e.g., fairness, justice). Our NLP model assessed prominence, frequency, and context, ensuring principled—not merely popular—value inclusion.
- **Instrumental Structures**: Auditability, explainability, and oversight are measured as enablers of trust, responsibility, and privacy. We do not treat these as standalone values but as functional indicators.
- **Dual-Role Values**: Innovation and sustainability were included only after showing consistent ethical significance, not just rhetorical repetition. This reflects our own standard of fairness in inclusion criteria.
- **Contextual Relevance**: Conditional values like privacy and autonomy were weighted based on relevance in sector-specific use cases. For instance, surveillance AI invoked higher thresholds for privacy and proportionality.

- **Accountability Clauses**: Only documents with explicit role assignments, update trails, and governance protocols met our traceability standard. This embeds our trust and responsibility values into evaluation, modeling "eating our own dogfood."
- **Human Rights Integration**: Privacy, agency, and non-discrimination were checked for alignment with UDHR (United Nations, 1948) and GDPR (European Union, 2016) anchors. These references were mapped using legal taggers[3] to ensure semantic and regulatory grounding.
- **Future-Orientation:** We required mechanisms for ethical drift detection, reassessment triggers, and value evolution. Ethical drift, where principled systems degrade under institutional normalization, has been documented in high-stakes fields such as law enforcement (Mann, 2020; Sternberg, 2012).

Recent scholarship on AI ethics further reinforces this need. Dai et al. (2022) demonstrate how user values shift over time in response to system interaction and technological embedding. De Cesare (2022) adds ontological clarity, showing that values themselves evolve structurally under new sociotechnical paradigms. Combined, these insights support our inclusion of adaptive value mechanisms and epistemic trust modeling—particularly for AI systems operating under conditions of uncertainty, transhuman integration, or deep sustainability.

In this spirit, the AI Moral Code project launches aimoralcode.org to continue iterating values through global input—modeling the same transparency and participatory ethics we ad.

[3] Legal taggers are natural language processing (NLP) tools or algorithms that identify and annotate legal concepts, statutes, and human rights principles within text, linking them to established legal frameworks such as the UDHR or GDPR. They ensure semantic alignment and regulatory traceability by mapping phrases to authoritative legal anchors or ontologies (e.g., European Legislation Identifier [ELI], LKIF, or LegalRuleML) (IBM, 2024; Bommarito II, Katz, & Detterman, 2018).

Methodological Details and Scoring Logic

Canonical Value Identification

The canonical value architecture was derived from convergence across over 291 ethics documents, not merely lexical clustering. Values were stratified into three operational tiers—Moral Core, Instrumental, and Conditional—based on simulation performance, ethical salience, and system-level deployment complexity. While institutional sources informed initial scaffolding, final inclusion required contextual stability, simulation resilience, and semantic coherence.

This stratification approach signals a shift from traditional value normalization—where significance is determined by frequency or cultural consensus—to performance-based ethical evaluation. In contrast to earlier reviews that relied on value frequency across documents to define ethical significance (Jobin, Ienca, & Vayena, 2019; Fjeld, Achten, Hilligoss, Nagy, & Srikumar, 2020; Floridi & Cowls, A unified framework of five principles for AI in society, 2019) our model emphasizes value performance within context-sensitive simulations. We do not assume that frequency implies ethical centrality—only that it signals discursive prominence. In our framework, a value earns its Moral Core status not from mention but from demonstrable resilience across complex, high-stakes simulations. Trust, for example, is not framed as a purely social construct (à la Aristotelian habit or Kantian duty), but as a system condition—an outcome of explainability, transparency, and behavioral predictability. In this way, the *AI Moral Code* redefines ethical centrality through system integrity rather than social convention, aligning moral valuation with the logic of machine-integrated co-agency.

Fox and DeMarco (2001) emphasize that moral reasoning in applied ethics must be derived from philosophical rigor, not polling or popularity. This contrasts with frequency-based ethics and aligns more closely with your stratified tier system based on virtue, duty, and consequence. Following Fox and DeMarco (2001), we resist treating value consensus or frequency as sufficient ethical justification. Instead, our methodology anchors values in structured moral reasoning—tested through consequence (simulation), duty (traceability), and virtue (semantic consistency)—consistent with the philosophical rigor expected in applied ethics.

Sectoral Representation and Semantic Sampling

Sectoral balance was preserved, but extraction focused on semantic matching rather than raw frequency. Sentence-BERT embeddings allowed precision-matching for ethical terms, while sectoral origin enabled post-hoc bias calibration. Semantic embedding was achieved using Sentence-BERT (Reimers & Gureyvch, 2019), allowing contextual value extraction at the sentence and paragraph level. This ensured rhetorical inflation (e.g., corporate repetition of 'transparency') did not artificially elevate prominence.

Sector Weighting

A Sector Weight Index (SWI) normalized value prominence by the normative authority of the issuing institution. Documents from governments and NGOs were given higher institutional weight, consistent with their roles in setting enforceable standards and ethical advocacy (Risse, Ruggie, & Zürn, 2013; Suchman, 1995; Bernstein, 2011). This approach ensures rhetorical volume does not mask moral significance.

Composite Value Score (CVS) LOGIC

Each value was scored using the following weighted formula:

$$CVS = \sum (TF^4 \times IDF^5 \times SWI^6 \times CM^7),$$

Where:

- TF x IDF anchors frequency context
- SWI calibrates sectoral legitmacy
- CM (Contextual Multiplier) boosts terms in high-impact zones (titles, principle clauses), and incorporates confidence scoring and simulation relevance.

Ethical Verification Framework

Each canonical value underwent triangulated validation via:

1. Val_Frequency_Counts for presence and semantic clarity
2. NRBC moral theory stratification (Virtue, Deontology, Consequentialism)
3. Scenario Simulation Performance, using the Val_Weighting_Schema to assess resilience in ethically complex environments (e.g., privacy under surveillance, fairness in algorithmic classification; autonomy in constrained systems).

Update on Canonical Value Evolution

Previous publications, including Hinrichs (2025), identified Justice, Transparency, Responsibility, Non-Maleficence, and Inclusivity as leading ethical values based on frequency and centrality. However, this deeper CVS-based derivation reveals a slightly evolved canonical structure that includes:

Core Moral Values

[4] TF = Term Frequency
[5] IDF = Inverse Document Frequency.
[6] SWI = Sector Weight Index
[7] CM = Contextual Multiplier

These are foundational ethical commitments, often deontological, virtue-based, or intrinsic to human moral frameworks. They form the non-negotiable moral core of the AI Moral Code.

1. **Beneficence** – Active promotion of good, well-being, and flourishing
2. **Dignity** – Recognition of intrinsic human worth and moral status
3. **Fairness** – Equity in treatment, distribution, and systemic outcomes
4. **Justice** – Procedural and distributive integrity across systems
5. **Responsibility** – Moral and institutional ownership of decisions
6. **Trust** – Foundational social bond enabling ethical co-functionality

Instrumental Values

These function as operational enablers. They help translate ethical commitments into system behavior, technical constraints, and measurable oversight mechanisms.

7. **Innovation** – Enables long-range adaptability and progress toward beneficial outcomes
8. **Sustainability** – Ensures long-term system resilience and environmental/social viability

Conditional Values

9. **Accountability** – Enables enforcement, traceability, and ethical audits, highly dependent on institutional capacity and legal culture.
10. **Autonomy** – Critical in domains involving consent, freedom, or human agency; varies in priority across medical, legal, and social domains.

11. **Inclusivity** – Accountability – Enables enforcement, traceability, and ethical audits, highly dependent on institutional capacity and legal culture.
12. **Privacy** – Context-sensitive value central to rights, identity, and power dynamics; especially critical in surveillance, biometric, and data-driven systems.

This expanded view integrates values such as Beneficence, Innovation, and Autonomy, reflecting their increased semantic prominence and policy emergence between 2022 and 2025.

Case Illustration: Sustainability

Composite Score: 1.07. A value such as Sustainability demonstrated stable but moderate frequency. However, sector-weighted contributions and contextual salience elevated their overall impact. Example: UNESCO's 2022 report AI for Climate Resilience in Emerging Economies triggered multiple context multipliers, contributing 0.092 to Sustainability's total CVS (UNESCO, 2022). This illustrates Sustainability's dual function—not only as a policy-anchored ethical imperative, but as an instrumental design principle essential for building adaptive, long-term AI infrastructures.

Interpretation: Sustainability does not dominate in raw frequency but exhibits strategic ethical relevance in infrastructure and long-term systems design.

Toward an Ethical Salience Tracker (EST)

To account for change over time, we introduce an Ethical Salience Tracker (EST)—a temporal recalibration model that continuously updates the relative weight of canonical values as the AI ethics landscape evolves. This framework recognizes that ethical priorities are not static; they shift in response to political climates, social movements, and emerging technologies.

The EST operates by monitoring three primary drivers:

- New legislation that elevates or constrains the prominence of specific values (e.g., privacy under GDPR, transparency under the EU AI Act)
- Semantic drift and emergent topics, such as the rising moral discourse around synthetic media, AI-generated personas, or digital sentience
- Sectoral discourse trends, which may amplify values like accountability in finance or fairness and trust in sports science—particularly as performance metrics, biometric monitoring, and team-based decision systems raise new ethical considerations in competitive and institutional settings

Each value is modeled like a moral signal, with ethical salience rising or falling over time. To capture this dynamism, we assign Ethical Volatility Metrics (EVMs) to each value, tracking fluctuations across documents, jurisdictions, and stakeholder communities. This structure allows AI governance to remain ethically grounded yet responsive recalibrating its core value set as societal priorities shift.

The EST model ensures that the AI Moral Code remains a living framework—empirically anchored, philosophically coherent, and adaptable to new ethical pressures without abandoning its foundational commitments.

Conclusion

This refined derivation of the AI Moral Code's canonical values offers a repeatable, data-grounded framework for ethical alignment. By situating the values within a stratified, sector-aware semantic scaffold—and validating them through philosophical and statistical rigor—this list forms the ethical DNA for value-centric AI. As AI systems increasingly govern high-stakes

environments, this refined canonical set provides a dynamic yet stable foundation for ethics-by-design. Future iterations of the AI Moral Code will continue to track semantic drift and contextual realignment, ensuring ethical fidelity in rapidly evolving domains.

Appendices (available upon request):

- Value Cognate Lists
- Raw TF-IDF Tables
- CVS Scripts (Python)
- Sectoral Normalization Models
- Volatility Snapshots (2018–2025)
- Old vs. New Canonical Rankings Table

References

Access Now. (2018, November). *Human Rights in the Age of Artificial Intelligence*. Retrieved from Access Now: https://www.accessnow.org/wp-content/uploads/2018/11/AI-and-Human-Rights.pdf

Bernstein, S. (2011). Legitimacy in intergovernmental and non-state global governance. *Review of International Political Economy, 18*(1), 17-51.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies.* Oxford University Press.

De Graff, G, Huberts, L., & Smulders, R. (2016). Coping with public value conflicts. *Administration & Society, 48*(9), 1101-1126.

European Commission. (2019, April 8). *Ethics Guidelines for Trustworthy AI*. Retrieved from European Commission: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council.* Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349-379. Retrieved from https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., . . . Vayena, E. (2018). AI4People--An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707. doi:https://doi.org/10.1007/s11023-018-9482-5

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines, 30*(3), 1-24. doi:https://doi.org/10.1007/s11023-020-09539-2

Pichai, S. (2018, June 7). *AI at Google: Our principles*. Retrieved from Google: https://blog.google/technology/ai/ai-principles/

Pontifical Academy for Life. (2020, February 28). *Rome Call for AI Ethics*. Retrieved from Pontifical Academy for Life: https://www.academyforlife.va/content/dam/pav/documenti%20pdf/2020/Call%20for%20AI%20Ethics/Rome_Call_for_AI_Ethics.pdf

Risse, T., Ruggie, J. G., & Zürn, M. (2013). The rational design of international institutions. *International Organization, 67*(4), 611-624.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Viking.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review, 20*(3), 571-610.

UNESCO. (2022). *AI for climate resilience in emerging economies.* Paris: UNESCO.

United Nations. (1948, December 10). *Universal Declaration of Human Rights*. Retrieved from https://www.un.org/en/about-us/universal-declaration-of-human-rights