Title: Deriving the 15 Canonical Values of the AI Moral Code: A Stratified Semantic Framework (2006–2025)

Author: Ran J. Hinrichs

Abstract:  
This paper outlines the empirical methodology used to derive the 15 canonical values of the AI Moral Code, based on a stratified semantic analysis of 291+ global AI ethics documents spanning from 2006 to 2025. The framework integrates frequency-weighted scoring, sectoral normalization, and contextual relevance to produce a ranked set of ethical values that serve as the normative core of value-aligned AI governance. This methodology establishes a replicable approach for future updates, enabling dynamic tracking of ethical priorities in the evolving AI ecosystem.

# Introduction

The AI Moral Code project aims to establish a universal yet adaptable ethical core for artificial intelligence, grounded in empirical evidence rather than abstract speculation. To ensure that the selected values are representative, actionable, and globally valid, this research draws from a corpus of 291+ AI ethics documents sourced from government agencies, intergovernmental bodies, academic institutions, corporations, NGOs, and religious organizations.

# Methodological Overview

2.1 Corpus Construction  
Documents were collected from 2006 to 2025 and classified by origin sector: Government, Industry, Academia, NGO, and Religious Organization. Each document was vectorized and ingested for analysis using NLP techniques.

2.2 Value Candidate Identification  
An initial pool of 43 candidate ethical values was identified using seed lists from IEEE, OECD, UNESCO, the European Commission, and prominent academic literature. Each value was defined by a core lemma and 3–7 high-frequency cognates (e.g., "justice" included fairness, equity, lawfulness).

2.3 Stratified Frequency Sampling  
Each document was coded for canonical value occurrence. To prevent sectoral overrepresentation, documents were stratified into equal weights across the five sectors. Term Frequency-Inverse Document Frequency (TF-IDF) was calculated for each term within 100–300 word context windows to distinguish signal from noise.

2.4 Sector Weighting  
A Sector Weight Index (SWI) was introduced to adjust for historical imbalance in document distribution. Government and NGO sectors were assigned higher weights due to policy-setting authority and grassroots moral insight, respectively.

# Value Scoring and Ranking

3.1 Composite Value Score (CVS)  
Each value was assigned a CVS calculated by:

CVS = ∑ (TF \* IDF \* SWI \* CM)

A close-up of a computer code

AI-generated content may be incorrect.

Where:

* TF = Term Frequency of the value and its cognates
* IDF = Inverse Document Frequency (global rarity adjustment)
* SWI = Sector Weight Index
* CM = Contextual Multiplier (boosted if term appeared in title, abstract, or principles section)

3.2 Relative Frequency Ranking  
Values were normalized on a 0–1 scale and ranked according to their composite scores. The top 15 were selected as canonical. For example:

* Justice = 2.21
* Transparency = 2.06
* Responsibility = 1.88
* ...
* Beneficence = 0.92

These scores were then proportionally scaled, making Justice the highest-weighted ethical concept across the corpus.

**Canonical Value Profile: Sustainability**

**Composite Score:** 1.07  
**Rank:** 10 out of 15 canonical values

Sustainability emerged as a consistently cited ethical value in AI governance literature from 2006 to 2025. The composite score of 1.07 reflects the weighted sum of Sustainability-related terms across 291 policy and ethics documents, using a stratified, sector-sensitive semantic model. The AI system evaluates each document by identifying:

* The presence of sustainability-related keywords (e.g., *resilience*, *renewal*, *long-term*)
* The location of those terms (e.g., title, abstract, principles section)
* The source sector of the document (e.g., NGO, Government, Academia), each weighted differently

**Example Document Analysis:** One representative file was UNESCO’s 2022 publication, *“AI for Climate Resilience in Emerging Economies.”* The AI flagged "resilience" in the **title**, “environmental sustainability” in the **abstract**, and found *Sustainability* explicitly listed in the **core principles section**. Because this document came from an intergovernmental body (sector weight: 1.3), appeared in prominent locations (contextual multiplier: 2.0), and used high-relevance language, it contributed a score of **0.092** toward Sustainability’s total of 1.07. This process repeated across the corpus, simulating an ESG-style index of ethical salience.

**Interpretation for Stakeholders:** Sustainability is ethically central for long-term governance, especially in climate-adjacent domains and future-proofing AI systems. Its score of 1.07 suggests it is not a dominant theme but a stable and high-relevance value, meriting consistent inclusion in AI policy development, risk assessment, and cross-sector standards.

**Recommended Use:** Include Sustainability as a key filter in scenario simulation, ethics audits, and intergenerational impact analysis. Especially critical when evaluating AI systems tied to infrastructure, resource management, or climate data prediction.

1. Canonical Values and Their Final Ranks
2. Justice
3. Transparency
4. Responsibility
5. Non-Maleficence
6. Inclusivity
7. Trust
8. Ethical Responsibility
9. Privacy
10. Innovation
11. Sustainability
12. Dignity
13. Collaboration
14. Autonomy
15. Human Rights
16. Beneficence
17. Toward a Dynamic Ethical Index

Given the volatility of AI policy literature, we propose an Ethical Volatility Metric (EVM) and Dynamic Ethical Index (DEI) that update value weights over time through continuous crawling and recalculation. Each canonical value becomes a "stock" tracked monthly, with volatility informed by:

* Emerging legislation
* Sectoral discourse shifts
* Semantic proximity to newly dominant themes (e.g., synthetic media, AGI alignment)

1. Conclusion

The derivation of the 15 canonical values from stratified semantic synthesis ensures that the AI Moral Code is empirically grounded, globally representative, and adaptable. This methodology provides a transparent, repeatable approach to ethical core construction and future refinement.

Appendices: Available upon request

* Cognate Lists per Canonical Value
* Raw TF-IDF tables
* Sectoral Normalization Weights
* Historical Volatility Snapshots (2018–2025)

### Python Script

Below is a conceptual Python script that mirrors the calculation methodology you specified for deriving Composite Value Scores (CVS) from your 291+ AI ethics documents. It follows your actual methodological steps: TF-IDF scoring, sectoral stratification, and contextual weighting.

This version assumes you have already preprocessed your document corpus and extracted term frequencies and inverse document frequencies per sector and term. The code uses placeholder data structures to illustrate the full pipeline.

import pandas as pd

import numpy as np

# Define canonical values and their cognates

canonical\_values = {

"Justice": ["justice", "fairness", "equity", "lawfulness"],

"Transparency": ["transparency", "openness", "disclosure", "explainability"],

"Responsibility": ["responsibility", "duty", "accountability", "ownership"],

"Non-Maleficence": ["non-maleficence", "do no harm", "caution", "safety"],

"Inclusivity": ["inclusivity", "inclusion", "participation", "access"],

"Trust": ["trust", "confidence", "faith", "assurance"],

"Ethical Responsibility": ["ethical responsibility", "moral duty", "ethical obligation"],

"Privacy": ["privacy", "confidentiality", "data protection"],

"Innovation": ["innovation", "creativity", "advancement", "progress"],

"Sustainability": ["sustainability", "resilience", "renewal", "eco"],

"Dignity": ["dignity", "respect", "worth", "sanctity"],

"Collaboration": ["collaboration", "cooperation", "shared action"],

"Autonomy": ["autonomy", "self-determination", "liberty"],

"Human Rights": ["human rights", "civil rights", "freedom"],

"Beneficence": ["beneficence", "altruism", "doing good"]

}

# Define sector weights (based on your normalization logic)

sector\_weights = {

"Government": 1.2,

"Industry": 0.8,

"Academia": 1.0,

"NGO": 1.3,

"Religious": 1.1

}

# Placeholder document-term matrix for demonstration

# Columns: term, tf (term frequency), idf (inverse doc freq), sector, in\_title\_or\_principles (bool)

df\_terms = pd.DataFrame({

"term": ["justice", "equity", "transparency", "trust", "responsibility", "privacy", "sustainability", "innovation"],

"tf": [120, 85, 130, 100, 95, 80, 70, 65],

"idf": [1.8, 2.1, 1.6, 1.5, 1.7, 1.9, 1.6, 1.5],

"sector": ["Government", "NGO", "Academia", "Industry", "Government", "NGO", "Religious", "Academia"],

"in\_title\_or\_principles": [True, False, True, False, True, False, False, False]

})

# Apply contextual multipliers and sector weights

def compute\_score(row):

CM = 1.5 if row["in\_title\_or\_principles"] else 1.0

SWI = sector\_weights[row["sector"]]

return row["tf"] \* row["idf"] \* SWI \* CM

df\_terms["score"] = df\_terms.apply(compute\_score, axis=1)

# Aggregate scores by canonical value

value\_scores = {}

for value, terms in canonical\_values.items():

matching\_scores = df\_terms[df\_terms["term"].isin(terms)]["score"]

value\_scores[value] = matching\_scores.sum() if not matching\_scores.empty else 0.0

# Normalize and sort

df\_scores = pd.DataFrame(list(value\_scores.items()), columns=["Canonical Value", "Composite Value Score (CVS)"])

df\_scores = df\_scores.sort\_values(by="Composite Value Score (CVS)", ascending=False).reset\_index(drop=True)

print(df\_scores)