# Appendix A: Canonical Value Derivation Methodology

# Appendix A: Canonical Value Derivation Methodology

## Overview

This appendix details the empirical and conceptual process by which the 15 Canonical Values of the AI Moral Code were derived. The process spanned three phases of iterative refinement and was grounded in a corpus of 291+ AI ethics documents published between 2006 and 2025.

## Phase 1–3: Corpus Construction and Value Extraction

The document set included government strategies (e.g., EU AI Act, U.S. AI Bill of Rights), international guidelines (UNESCO, OECD), corporate ethics statements (IBM, Google, Microsoft), academic inventories (Harvard, Oxford), and landmark meta-reviews (Jobin et al., Hinrichs 2024, Bonnici 2023).

Each document was:

- Indexed and assigned a document ID

- Sector-tagged (government, NGO, industry, academia)

- Analyzed using semantic keyword extraction and token frequency modeling

This resulted in an initial lexicon of 150+ ethical terms.

## Refinement Phases: From Candidates to Canonical Values

Three iterative review rounds were conducted using frequency tables and semantic clustering. Redundant terms were collapsed, synonyms merged, and values reweighted based on:

- Raw frequency (token count across documents)

- Cross-sector recurrence

- Cultural universality

- Independent moral function

\*\*Fairness\*\*, although frequent, was \*\*excluded\*\* from the final canonical set due to:

1. Semantic overload (different meanings across sectors and cultures)

2. Subsumability under Justice, Inclusivity, Non-Maleficence, and Responsibility

3. Poor operational distinctiveness in AI systems design

## Final Canonical Values (n = 15)

The following values emerged with the highest empirical and conceptual integrity:

1. Trust

2. Transparency

3. Responsibility

4. Non-Maleficence

5. Autonomy

6. Inclusivity

7. Justice

8. Privacy

9. Sustainability

10. Dignity

11. Innovation

12. Collaboration

13. Human Rights

14. Beneficence

15. Ethical Responsibility

Each value was evaluated for structural independence, cross-sector compatibility, and suitability for AI ethics enforcement and architectural modeling.

## Validation

These canonical values were validated against:

- Jobin et al., \*Nature Machine Intelligence\*, 2019

- Harvard Berkman Klein Ethics Inventory

- Hinrichs Meta-Frequency Analysis (2024)

- Floridi’s foundational ethics frameworks

- Legal domain reviews (Bonnici 2023)

- Comparative token analysis from 291+ global documents

They form the ethical foundation of the AI Moral Code framework and its NRBC (Normative, Regulatory, Behavioral, and Conceptual) layers.