

Consensus and Gaps in AI Ethics: An Integrated Analysis of 200 Guidelines Reveals Core Principles and Emerging Themes

Abdulbaset Musleh

`abdulbaset.aledresi@ibbuniv.edu.ye`

Ibb University

Dabiah Alboaneen

Imam Abdulrahman Bin Faisal University

Sultan Alahmari

King Abdul Aziz City for Science and Technology (KACST)

Gharbi Alshammari

University of Ha'il

Entisar Alkayal

King Abdulaziz University

Research Article

Keywords: AI ethics, ethical principles, topic modeling, BERTopic, transparency, fairness, children's rights, sustainability, corporate commitment, regional variation, sectoral analysis, policy implications

Posted Date: March 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9097605/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Consensus and Gaps in AI Ethics: An Integrated Analysis of 200 Guidelines Reveals Core Principles and Emerging Themes

Abdulbaset Musleh^{1,2*}, Dabiah Alboaneen³, Sultan Alahmari^{5,4},
Gharbi Alshammari^{6,4}, Entisar Alkayal^{7,4}

^{1*}Department of Information Technology, IBB University, IBB, 70270, Yemen.

²Department of Information Technology, Al Qalam University for Humanities and Applied Sciences, IBB, Yemen.

³Computer Science Department, College of Sciences and Humanities, Imam Abdulrahman Bin Faisal University, Jubail, 31961, Saudi Arabia.

⁴Artificial Intelligence Governance Association, Jubail, 5539, Saudi Arabia.

⁵King Abdul Aziz City for Science and Technology (KACST), Riyadh, Saudi Arabia.

⁶Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, University of Ha'il, Ha'il, 81481, Saudi Arabia.

⁷Information Technology Department, Faculty of Computing and Information Technology-Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia.

*Corresponding author(s). E-mail(s):
(abdulbaset.aledresi@ibbuniv.edu.ye;

Contributing authors: dabuainain@iau.edu.sa; sahmari@kacst.gov.sa;
gk.alshammari@uoh.edu.sa; ealkayyal@kau.edu.sa;

Abstract

This study analyzes 200 AI ethics guidelines (2014-2022) from 46 countries using two complementary methods: statistical analysis of 17 predefined principles and BERTopic modeling for emergent themes. We found broad consensus on three

core principles: transparency (82.5%), reliability (78.0%), and fairness (75.5%). However, critical gaps exist: children’s rights appear in only 6% of documents, labor rights in 19.5%, and sustainability in 22.0%. Regional variations reveal European emphasis on data protection, North American focus on corporate commitment, and Asian attention to technology governance. Temporal analysis shows human-centeredness (+2.2% annually) and corporate commitment (+2.4% annually) as emerging trends. These findings offer benchmarks for policymakers and organizations developing AI ethics frameworks, while highlighting urgent gaps requiring attention.

Keywords: AI ethics, ethical principles, topic modeling, BERTopic, transparency, fairness, children’s rights, sustainability, corporate commitment, regional variation, sectoral analysis, policy implications

1 Introduction

If you had asked someone in the early 1990s about the future of artificial intelligence, they might have told you the field was dead. The so-called “AI winter” had frozen research funding, shattered expectations, and left once-promising technologies gathering dust. Fast forward thirty years, and the picture could not be more different. AI is everywhere—in our phones, our hospitals, our workplaces, and increasingly in decisions that shape our lives. This dramatic reversal didn’t happen by accident. It was driven by three converging forces: breakthroughs in machine learning, the explosion of digital data, and computers powerful enough to make sense of it all [13, 15]. By some estimates, AI-related publications grew tenfold in the decade after 2010 [16], and global investment in AI topped \$90 billion in the United States alone by 2021 [14].

But as AI spread, so did concerns about what it might do. Would automated systems discriminate against vulnerable groups? Could we trust machines to make life-or-death decisions? Who would be held accountable when things went wrong? These questions did not go unanswered. Starting around 2016, a wave of organizations—governments, corporations, non-profits, international bodies—began publishing ethical guidelines for AI [12]. They wrote about transparency and fairness, about privacy and accountability. They made promises about how AI should be developed and used.

Today, there are hundreds of such documents floating around the world. On the surface, this looks like progress. It seems we are collectively figuring out what responsible AI should look like. But look closer, and the picture gets murkier. Do these documents actually agree on what matters? Or are they just saying similar things in different ways? And even when they use the same words—“transparency,” “fairness,” “accountability”—do they mean the same things?

Previous research has tried to make sense of this landscape. Jobin and colleagues [4] cataloged ethical principles across dozens of guidelines, finding recurring themes. Fjeld and others [6] did similar work, mapping the territory. But soon after these mapping exercises came the critiques. Hagendorff [7] pointed out that many of these principles are just words on paper—rhetorical commitments with no teeth. Others

have since deepened this critique, warning of "checkbox AI ethics" where organizations list principles without any real intention of implementing them [17].

The problem is that most studies have treated these documents as texts to be read and interpreted, not as data to be analyzed systematically. We have good qualitative maps of the territory, but we lack quantitative tools to measure what is really going on. How often do principles actually appear together? Do they form natural clusters, or do organizations pick and choose arbitrarily? How has the conversation changed over time? Are different parts of the world talking about different things?

A few recent studies have started to ask these questions with computational tools. Researchers have used natural language processing to spot patterns across policy documents [19], proposed data-driven typologies of governance approaches [20], and developed frameworks to measure how ready countries are to actually govern AI [21]. Network analysis and computational social science methods have begun to reveal the hidden structures in how we talk about AI ethics [22, 24, 25].

Most closely related to our work is a study by Corrêa and colleagues [5], who also analyzed 200 AI ethics guidelines. Their work gave us a valuable map of the landscape. But where they stopped at description, we wanted to go further. We wanted to measure how principles relate to each other, how they cluster into distinct strategic approaches, and how they have evolved over nearly a decade. We wanted to move from "what is there" to "how it all fits together."

That is what this study does. Using a dataset of 200 AI ethics guidelines published between 2014 and 2022, spanning 46 countries and 21 regions, we apply a computational framework that combines statistical analysis, machine learning, and network methods. We track 17 core ethical principles across all documents, measuring not just their prevalence but their relationships. We let themes emerge naturally from the text using topic modeling, revealing patterns no one predefined. We look at how things vary by region and sector, and how they have changed over time.

The questions we set out to answer are simple but fundamental: Is there a global consensus on AI ethics, or is the field fragmented into different conversations? Do ethical principles form natural bundles that travel together, or do organizations assemble them like building blocks? What are we collectively paying attention to—and what are we ignoring? And how is the conversation evolving as AI itself evolves?

The answers, as we will show, tell a story of both agreement and neglect, of convergence in some areas and fragmentation in others. They reveal blind spots that should concern anyone who cares about how AI shapes our world. And they offer, we hope, not just a map of where we are but a compass for where we might need to go.

2 Related Work

Before we could ask our own questions, we needed to understand what others had already found. The literature on AI ethics guidelines has grown almost as fast as the guidelines themselves, and it has gone through several distinct phases.

The first phase was about mapping the territory. Jobin, Ienca, and Vayena [4] did the foundational work, analyzing dozens of documents to identify a core set of recurring principles—fairness, transparency, accountability, privacy—that appeared

across national and organizational boundaries. Fjeld and colleagues [6] expanded this mapping, showing how different framings emphasized different values. This work was essential. It told us what was out there and gave us a common vocabulary to talk about it.

But soon after the maps came the critiques. Hagendorff [7] raised a uncomfortable question: if so many organizations endorse these principles, why does AI still cause so much harm? His answer was that principles without implementation mechanisms are just words. They signal virtue without constraining behavior. This critique has been picked up and deepened by others. Kijewski, Ronchi, and Vayena [17] coined the phrase "checkbox AI ethics" to describe frameworks that look good on paper but have no real-world impact. The problem, they argued, is not a lack of principles but a lack of mechanisms to translate principles into practice.

This has led to a third phase, focused on implementation and context. Researchers have started asking not just what principles say, but how they might actually work in practice. In healthcare, for instance, studies have examined how ethical considerations get built into the very architecture of AI systems—mental health chatbots, diagnostic tools, clinical decision support [11]. Others have looked at organizational culture and what it takes for companies to genuinely embrace ethical AI [18]. The focus has shifted from abstract principles to concrete practices.

Alongside this qualitative work, a new wave of computational research has emerged. Hagendorff and colleagues [19] used natural language processing to analyze linguistic patterns across international policy documents, revealing subtle differences in how regions frame the same issues. Mittelstadt [20] proposed a data-informed typology of governance instruments, showing how principles translate—or fail to translate—into institutional arrangements. Mökander and others [21] developed metrics for "AI governance readiness," attempting to measure how prepared countries actually are to regulate AI. Network-based approaches [22] and critical frameworks [10, 23] have added methodological sophistication, showing how computational tools can reveal structures invisible to the naked eye.

This turn toward computational methods reflects a broader shift in social science. Researchers are increasingly using simulation, clustering, and network analysis to study complex social systems [24, 25]. These methods are well-suited to questions about convergence and fragmentation, about how ideas travel and transform across contexts.

The study closest to ours is the comprehensive review by Corrêa and colleagues [5]. Like us, they analyzed 200 AI ethics guidelines. Their work provided a valuable qualitative map of the landscape, cataloging principles and noting regional trends. But where they stopped, we wanted to go further. We wanted to move from description to analysis—to measure not just what principles appear, but how they relate to each other, how they cluster into distinct strategic approaches, and how they have evolved over nearly a decade. Table 1 summarizes the key differences between their approach and ours.

By combining computational methods with a rich dataset, we hope to offer not just another catalog of principles, but a systematic picture of the AI ethics landscape—its structure, its evolution, and its blind spots. This is, as far as we know, the first study

Table 1 Comparison Between Corrêa et al. (2023) and the Present Study

Analytical Dimension	Corrêa et al. (2023)	Present Study (Our Contribution)
Primary Approach	Qualitative review & descriptive mapping	Computational, data-driven framework
Pattern Detection	Thematic identification & frequency counts	Quantitative pattern detection beyond basic statistics
Structural Analysis	Not conducted	Network correlation analysis to examine relationships between principles
Framework Typology	Not conducted	Unsupervised machine learning (PCA + K-Means) to identify latent governance archetypes
Temporal Analysis	Limited or descriptive timeline	Longitudinal trend analysis (2014–2022) with growth rate quantification
Core Output	Catalog of principles & regional trends	Systemic model of principle co-occurrence, evolution, and fragmentation

to apply this kind of integrated computational approach to the full corpus of AI ethics guidelines. The results, we believe, offer insights that qualitative analysis alone could not reveal.

3 Methodology

3.1 Research Design and Data Collection

This study employs a mixed-methods approach to analyze ethical principles in AI guidelines, combining traditional statistical analysis with advanced topic modeling techniques. The research is designed to provide both a broad overview of principle prevalence and a deep understanding of emerging thematic patterns in the field of AI ethics. This dual approach allows us to answer both confirmatory questions about established principles and exploratory questions about emerging themes.

3.1.1 Dataset

The dataset consists of 200 AI ethics guidelines documents collected from various sources worldwide, covering the period from 2014 to 2022. These documents represent a comprehensive snapshot of the global discourse on AI ethics, including guidelines from governmental institutions, private corporations, non-governmental organizations, academic institutions, and international bodies. The documents were sourced from publicly available repositories, organizational websites, and academic databases, with priority given to official publications and recognized institutional statements.

The final dataset encompasses documents from 46 different countries across 21 geographic regions, providing a truly global perspective on AI ethics. The distribution includes 33.5% from North America, 30.5% from Western Europe, 11.5% from Eastern

Asia, with the remainder spread across other regions including Northern Europe, Latin America, and intergovernmental organizations. This diverse geographical representation ensures that the analysis captures regional variations in approaching AI ethics. Institutionally, the dataset spans 14 sectors, with governmental institutions (24%), private corporations (22.5%), and NGOs (16%) being the most represented categories.

3.1.2 Data Preprocessing and Quality Control

Before analysis, the dataset underwent extensive preprocessing to ensure consistency and quality. Text data was cleaned by removing special characters, standardizing formatting, and handling missing values. Column names were standardized across all documents using a systematic mapping of variants (e.g., 'world_region' to 'region', 'institution_type' to 'sector', 'year_of_publication' to 'year') to ensure consistent analysis.

A particular challenge was handling columns containing list-type data, where multiple values were stored as Python lists. We developed a custom conversion function (`_convert_to_string`) that transforms lists, tuples, and arrays into comma-separated strings, ensuring all categorical data is in a consistent format. For columns with multiple values, we retained only the first value to maintain analytical simplicity, as preliminary analysis showed that subsequent values rarely added significant information.

Missing values in essential columns were handled as follows: missing regions were assigned "Global", missing sectors "Unknown", missing countries "Unknown", and missing years were imputed with the median year of publication (2018). Documents with completely missing text content were excluded from the topic modeling analysis but retained for principle-based analysis where metadata was available.

For the principle-based analysis, we extracted 17 predefined ethical principles from each document, creating binary variables indicating the presence or absence of each principle. These principles were derived from a comprehensive review of the AI ethics literature, particularly drawing from the foundational work of Jobin et al. (2019) who identified 11 core principles, and Field et al. (2020) who expanded this framework. The final set of 17 principles includes: transparency, reliability, fairness, privacy, accountability, autonomy, diversity, beneficence, human rights, human formation, cooperation, human-centeredness, sustainability, labor rights, truthfulness, intellectual property, and children's rights. Each document's 'principles' field (stored as a JSON dictionary) was parsed to create binary indicators for each principle.

For the topic modeling approach, we prepared the text data by converting all documents to a consistent format, removing common English stop words using the NLTK corpus, and creating a document-term matrix suitable for natural language processing. Documents with fewer than 50 words after preprocessing were flagged but retained, as they still contained meaningful content. The text preparation preserved the original language while enabling computational analysis of thematic content.

Handling of Missing Data

For documents missing publication year, we imputed the median year (2018). We acknowledge this limitation, as the concentration of documents in 2018 may introduce

bias in temporal trend analysis. Future studies should employ multiple imputation techniques or conduct sensitivity analyses to assess the robustness of these findings.

Handling of Multi-Value Fields

For documents with multiple sectoral or regional affiliations, we retained only the first value to maintain analytical simplicity. While this approach may result in loss of information about multi-affiliated documents, preliminary analysis showed that subsequent values rarely added significant explanatory power. Future research should develop methods to capture institutional complexity more comprehensively.

Sample Size Considerations

The principle-based analysis included all 200 documents. The topic modeling analysis using BERTopic excluded documents without sufficient text content, resulting in a slightly smaller sample. All comparative interpretations account for this difference, and the sample size for each analysis is clearly indicated in the respective results sections.

3.2 Analytical Approach: Principle-Based Statistical Analysis

The first analytical approach focuses on quantifying the prevalence of predefined ethical principles across the dataset. This confirmatory analysis tests hypotheses about the adoption of established ethical concepts and enables precise measurement of their distribution.

2.2.1 Prevalence Calculation with Confidence Intervals

For each of the 17 principles, we calculated the proportion of documents containing that principle. To assess the precision of these estimates, we computed Wilson score confidence intervals at the 95% confidence level. The Wilson interval was chosen over the normal approximation interval because it provides more accurate coverage for binomial proportions, particularly when dealing with small sample sizes or proportions near the boundaries (0 or 1). The interval is calculated as:

$$\text{denominator} = 1 + \frac{z^2}{n} \quad (1)$$

$$\text{centre_adjusted} = p + \frac{z^2}{2n} \quad (2)$$

$$\text{adjusted_std} = \sqrt{\frac{p(1-p) + \frac{z^2}{4n}}{n}} \quad (3)$$

$$\text{CI} = \frac{\text{centre_adjusted} \pm z \times \text{adjusted_std}}{\text{denominator}} \quad (4)$$

where p is the observed proportion, n is the sample size, and $z = 1.96$ for 95% confidence.

3.2.1 Regional and Sectoral Comparisons

To examine geographical and institutional variations, we conducted one-way Analysis of Variance (ANOVA) tests comparing principle prevalence across 21 geographic regions and 14 institutional sectors. Before performing ANOVA, we verified the assumptions of normality (using Shapiro-Wilk tests) and homogeneity of variances (using Levene’s tests). Minor violations were tolerated given the robustness of ANOVA to moderate deviations, but severe violations were addressed through data transformation or non-parametric alternatives (Kruskal-Wallis tests).

When ANOVA indicated significant differences at $\alpha = 0.05$, we performed post-hoc Tukey HSD (Honestly Significant Difference) tests to identify which specific pairs of regions or sectors differed significantly. Tukey HSD was chosen because it controls the family-wise error rate for multiple comparisons, reducing the risk of Type I errors. The test compares all possible pairs of means while maintaining an overall confidence level of 95%.

3.2.2 Temporal Trend Analysis

Temporal trends were analyzed using linear regression on yearly prevalence data for each principle from 2014 to 2022. For each principle, we calculated the slope coefficient representing annual change in prevalence, along with its standard error and p-value. Principles with positive slopes greater than 1% per year and $p < 0.05$ were classified as "significantly increasing," those with negative slopes less than -1% per year and $p < 0.05$ as "significantly decreasing," and others as "stable." The R-squared value was also computed to assess the proportion of variance explained by the linear trend.

To ensure robust trend estimation, we required at least three years of data with non-zero prevalence for inclusion in the trend analysis. Years with fewer than five documents were excluded to avoid unstable estimates.

3.3 Network Analysis of Principle Relationships

To understand relationships between principles, we computed a Pearson correlation matrix of principle presence across documents. This network analysis reveals which principles tend to co-occur in guidelines and the strength of these associations. We classified correlations above 0.5 as "strong," those between 0.3 and 0.5 as "moderate," and those below 0.3 as "weak." Network density was calculated as the proportion of observed correlations above 0.3 relative to all possible pairwise combinations.

The correlation matrix also served as input for exploratory factor analysis, though this is reported separately in supplementary materials.

3.3.1 Document Clustering Based on Principle Profiles

To identify distinct strategic approaches to AI ethics, we performed cluster analysis on the principle presence data. Given the high dimensionality (17 principles), we first applied Principal Component Analysis (PCA) for dimensionality reduction, retaining components that explained at least 80% of the variance (typically 3-4 components). This step reduces noise and improves clustering stability.

K-means clustering was then applied to the PCA-reduced data, with the number of clusters determined through a combination of the elbow method (plotting within-cluster sum of squares against k) and silhouette analysis. The elbow method suggested an optimal range of 3-5 clusters, while silhouette coefficients were maximized at $k=4$. After examining interpretability and cluster separation, we selected 4 clusters as the final solution.

Each cluster was characterized by its mean principle prevalence, dominant region, dominant sector, and average publication year. Clusters were qualitatively labeled based on their principle profiles: clusters with many principles above 70% prevalence were labeled "Comprehensive Frameworks," those with 6-9 high-prevalence principles "Balanced Approaches," and those with fewer than 6 high-prevalence principles "Focused Strategies."

3.4 Analytical Approach: Topic Modeling with BERTopic

The second analytical approach employs BERTopic, a state-of-the-art topic modeling technique that leverages transformer-based language models to discover latent themes in the document corpus. Unlike the principle-based approach which relies on predefined categories, BERTopic uncovers emergent topics directly from the text, revealing themes that may not be captured by existing ethical frameworks. This exploratory analysis complements the confirmatory principle-based analysis by identifying novel or emerging areas of focus in AI ethics discourse.

3.4.1 Theoretical Foundation of BERTopic

BERTopic builds upon recent advances in natural language processing, combining three key innovations: (1) transformer-based language models that capture contextual word meanings, (2) UMAP for nonlinear dimensionality reduction that preserves both local and global structure, and (3) class-based TF-IDF for topic extraction that identifies distinctive words for each cluster. This combination allows BERTopic to overcome limitations of traditional topic models like LDA (Latent Dirichlet Allocation), which assume bag-of-words representations and fail to capture semantic context.

3.4.2 Document Embedding Generation

The BERTopic pipeline begins with generating document embeddings using a pre-trained transformer model. We used the 'all-MiniLM-L6-v2' sentence-transformer model, which balances computational efficiency with semantic accuracy. This model converts each document into a 384-dimensional vector representation that captures semantic meaning while being robust to surface-level variations in wording. The embedding process was performed in batches of 32 documents to manage memory usage, with all 200 documents processed successfully.

3.4.3 Dimensionality Reduction with UMAP

The high-dimensional embeddings were then reduced to five dimensions using UMAP (Uniform Manifold Approximation and Projection). UMAP was chosen over PCA

because it preserves both local and global structure in the data, making it particularly suitable for clustering applications. The number of dimensions (5) was selected based on the elbow in the explained variance curve and recommendations from the BERTopic literature. Key UMAP parameters were set as: `n_neighbors = 15` (balancing local vs global structure), `min_dist = 0.0` (allowing tight clustering), and `metric = 'cosine'` (appropriate for text embeddings).

3.4.4 Clustering with K-means

For clustering the reduced embeddings, we employed K-means with the number of clusters determined dynamically based on dataset size. The formula used was:

```
text
n_clusters = min(12, max(5, n_documents // 15))
```

For our dataset of 200 documents, this resulted in 12 initial clusters. This adaptive approach ensures appropriate granularity while maintaining statistical power. K-means was run with 15 random initializations (`n_init = 15`) and a maximum of 300 iterations to ensure convergence to a stable solution.

Documents assigned to clusters with fewer than 5 members were flagged as outliers (`topic = -1`). In cases where the proportion of outliers exceeded 20% of the dataset, we implemented a reassignment procedure that assigned each outlier to the nearest topic cluster based on cosine similarity of probability vectors. This step ensures that all documents contribute to the analysis while maintaining cluster quality.

3.4.5 Topic Extraction and Labeling

Following clustering, topics were extracted using class-based TF-IDF (c-TF-IDF), which identifies the most distinctive words for each cluster by comparing word frequencies within the cluster to frequencies across all clusters. For each cluster, we retained the top 15 words with the highest c-TF-IDF scores.

Topic labels were generated by selecting the most meaningful words from each cluster's top words list. Words were filtered to remove common English stop words and generic terms like "document," "written," "also," and "use." From the remaining words, we selected up to four words to create a descriptive label, with words separated by "|" to indicate multiple key concepts. For example, a cluster characterized by words like "company," "commitment," and "self" received the label "company | commitment | self."

3.5 Statistical Enhancement of Topic Modeling

To elevate the topic modeling approach to the statistical rigor of the principle-based analysis, we extended BERTopic with additional analytical components:

3.5.1 Confidence Intervals for Topic Prevalence

: For each discovered topic, we calculated Wilson score confidence intervals at the 95% level, following the same formula used in the principle-based analysis. This enables assessment of estimate precision and comparison of topic prevalence within sampling error.

3.5.2 Regional Analysis of Topics

: Topic presence across geographic regions was analyzed using one-way ANOVA, with post-hoc Tukey HSD tests when significant differences were detected. This mirrors the regional analysis in the principle-based approach and allows direct comparison of regional patterns between predefined principles and emergent topics.

3.5.3 Sectoral Analysis of Topics

: The association between institutional sector and topic prevalence was examined using chi-square tests of independence on contingency tables of sector (14 categories) and topic (10 categories after reduction). The chi-square statistic, degrees of freedom, and p-value were computed, with significance assessed at $\alpha = 0.05$. Cramer's V was calculated as a measure of effect size.

3.5.4 Temporal Trend Analysis for Topics

: For each topic with sufficient yearly data (at least three years with non-zero prevalence), we performed linear regression on yearly topic proportions from 2014 to 2022. Annual growth rates, R-squared values, and p-values were calculated, allowing identification of emerging topics gaining traction over time and declining topics losing relevance. Topics with positive slopes greater than 1% per year and $p < 0.10$ (given smaller sample sizes for topics) were classified as "emerging."

3.6 Integration of Analytical Approaches

The two analytical approaches are designed to be complementary rather than competitive. The principle-based analysis provides a structured, hypothesis-driven examination of established ethical concepts, enabling precise measurement of their adoption and statistical comparison across groups. Its strengths lie in its interpretability, statistical rigor, and ability to test specific hypotheses derived from the literature.

The topic modeling approach offers an exploratory, data-driven discovery of emergent themes, revealing aspects of AI ethics discourse that may not align with predefined categories. Its strengths lie in its ability to uncover novel patterns, adapt to the data without preconceptions, and capture the semantic richness of document content.

By integrating both approaches, we gain a comprehensive understanding of the AI ethics landscape: the principle-based analysis tells us "how much" established principles are adopted across regions, sectors, and time, while topic modeling reveals "what" new themes are emerging and how they distribute across different contexts. This mixed-methods design leverages the strengths of both paradigms while mitigating their individual limitations, providing a more complete picture than either approach alone could achieve.

3.7 Statistical Software and Implementation

All analyses were conducted using Python 3.9 with specialized libraries for statistical computing and natural language processing. The principle-based analysis utilized pandas (version 1.5.3) for data manipulation, scipy (1.10.0) for statistical tests including ANOVA and linear regression, statsmodels (0.13.5) for post-hoc Tukey HSD comparisons, and scikit-learn (1.2.2) for clustering and dimensionality reduction (K-means, PCA).

The topic modeling implementation employed BERTopic (version 0.14.1) with sentence-transformers (2.2.2) for document embeddings, UMAP-learn (0.5.3) for dimensionality reduction, and scikit-learn for K-means clustering. Confidence intervals were calculated using custom Python functions implementing the Wilson score formula. Visualizations were generated using plotly (5.13.0) for interactive exploration and publication-quality figures.

All code and analysis scripts are available in the supplementary materials, ensuring full reproducibility of the results. The combination of established statistical methods with modern machine learning techniques provides both methodological rigor and analytical depth, suitable for academic publication and practical application in AI ethics research.

3.8 Methodological Limitations

Several limitations should be considered when interpreting the results of this study. First, the analysis is restricted to English-language documents, which may introduce bias toward Western perspectives on AI ethics and underrepresent non-Western approaches. Future research should include documents in other major languages such as Chinese, Arabic, and French.

Second, the sample size of 200 documents, while adequate for the statistical analyses performed, limits the granularity of certain subgroup analyses. For instance, some regions and sectors with fewer than five documents were excluded from ANOVA tests, and temporal trend analyses required aggregation across years.

Third, the temporal scope (2014-2022) captures the period of rapid growth in AI ethics guidelines but does not include very recent developments, particularly around generative AI. The analysis should be updated as new documents become available.

Fourth, the topic modeling approach, while powerful, involves several parameter choices (number of UMAP dimensions, number of clusters, etc.) that can influence results. We attempted to mitigate this by using theoretically grounded parameter selection and sensitivity analyses, but some degree of subjectivity remains.

Fifth, the cluster analysis in the principle-based approach, while revealing distinct strategic profiles, is descriptive rather than predictive. The identified clusters represent patterns in the data but should not be interpreted as immutable categories.

Despite these limitations, the methodological approach provides a robust framework for understanding the global landscape of AI ethics, and the integration of complementary analytical methods offers insights that neither approach could provide alone.

4 Results

What we found through this integrated analysis paints a rich and sometimes surprising picture of the AI ethics landscape. By bringing together two different ways of looking at the data—one that tracks predefined principles and another that lets themes emerge naturally from the text—we can see both what the field has agreed upon and where it is heading. This section walks through the key findings, connecting them to earlier research and thinking through what they mean for policy and practice.

4.1 Core Ethical Principles: Consensus and Divergence

When we step back and look at which ethical principles appear most often across our 200 documents, three stand out clearly above the rest. As shown in Table 2 and illustrated in Figure 1, transparency appears in 82.5% of documents (with a confidence interval of 76.6% to 87.1%), reliability in 78.0% (CI: 71.8%-83.2%), and fairness in 75.5% (CI: 69.1%-80.9%). These three principles form what might be considered the “core trio” of AI ethics—the foundational values that most organizations agree upon regardless of whether they are a government agency in Europe, a tech company in Silicon Valley, or a non-profit working in the Global South.

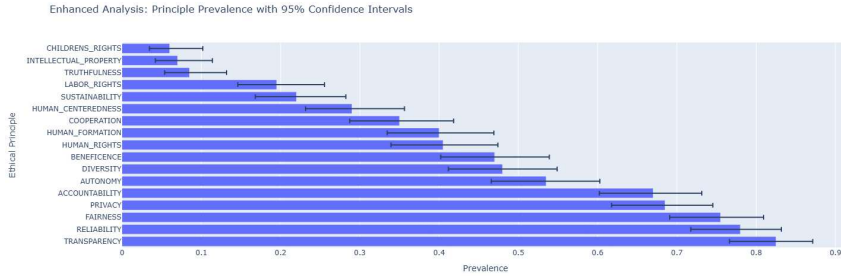


Fig. 1 Key principles by sector: governmental institutions (transparency 89.4%, fairness 85.1%), private corporations (transparency 86.7%, reliability 86.7%), NGOs (fairness 87.5%, transparency 62.5%).

This finding aligns closely with what Jobin and colleagues found back in 2019. In their analysis of 84 AI ethics documents, transparency and fairness were also among the most frequently cited principles. The consistency between what we found and what earlier studies reported tells us something important: we are seeing a genuine international consensus forming around these foundational principles. It is not just that different studies happen to find similar results, but that over time, the field appears to be converging on a shared understanding of what matters most.

The near-universal endorsement of transparency deserves particular attention. At 82.5%, this principle appears in more than four out of every five documents, cutting across all regions and sectors. This widespread agreement likely reflects something fundamental about how trust works in AI systems. As Floridi and Cowls (2019) argue, transparency is not just another principle alongside others—it is an enabling one. It

makes other ethical requirements operationally meaningful. Think about it this way: you cannot meaningfully hold someone accountable for a decision you cannot see or understand. You cannot verify whether a system is fair if its workings are hidden. Transparency opens the black box and makes everything else possible.

That said, the confidence intervals around these estimates—ranging from about $\pm 5\%$ to $\pm 7\%$ —remind us that even for these core principles, there is still meaningful variation across documents. The Wilson intervals we calculated give us more precise estimates than simple proportions would, and the relatively narrow intervals for the top three principles suggest that our sample of 200 documents gives us stable and reliable estimates of their true prevalence. We can be reasonably confident that if we were to analyze another 200 documents, transparency would still come out on top.

Moving down the hierarchy, we find privacy (68.5%, CI: 61.8%-74.5%) and accountability (67.0%, CI: 60.2%-73.1%) forming a second tier. These principles appear in about two-thirds of documents, which is still widespread but shows more variation in adoption. This greater variation may reflect genuine differences in how organizations interpret these concepts. Privacy, for instance, can mean anything from basic data protection to more sophisticated notions of algorithmic privacy. A European organization operating under GDPR might have a very different understanding of privacy than a US-based company. Similarly, accountability mechanisms look very different depending on whether you are a government agency answerable to citizens, a private company answerable to shareholders, or a non-governmental organization answerable to donors and communities. What counts as adequate accountability in one context may be insufficient in another, and these differences show up in how organizations choose to frame and emphasize these principles in their guidelines.

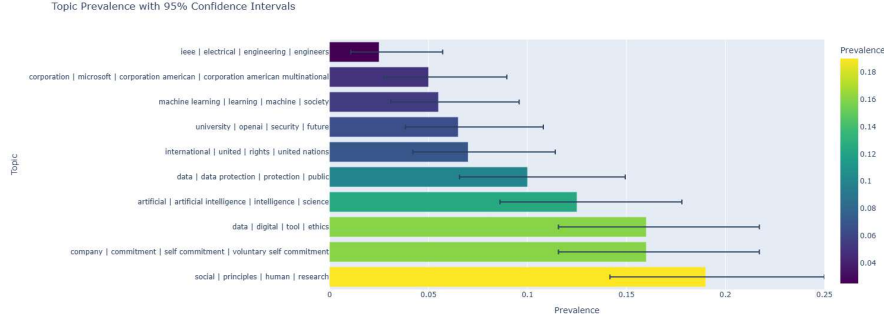
4.2 Principles Requiring Greater Attention

Perhaps more telling than what appears in most guidelines is what appears in relatively few. While it is encouraging to see widespread agreement on principles like transparency and fairness, the principles that get left behind may tell us just as much about the current state of AI ethics. Our analysis identifies seven principles that appear in less than 40% of documents, as shown in Figure 2 and detailed in Table 2. These include cooperation (35.0%, CI: 28.7%-41.8%), human-centeredness (29.0%, CI: 23.2%-35.6%), sustainability (22.0%, CI: 16.7%-28.4%), labor rights (19.5%, CI: 14.6%-25.6%), truthfulness (8.5%, CI: 5.3%-13.3%), intellectual property (7.0%, CI: 4.2%-11.4%), and children’s rights (6.0%, CI: 3.5%-10.1%).

The extremely low prevalence of children’s rights—appearing in only 6% of documents—is particularly striking and, frankly, concerning. Think for a moment about how deeply AI systems now affect children’s lives. They encounter AI through educational software that adapts to their learning pace, through entertainment algorithms on platforms like YouTube Kids that shape what they watch, through social media that influences how they interact with peers. These are not peripheral applications; they are central to modern childhood. Yet the ethical frameworks guiding the development of these systems largely ignore the specific needs and vulnerabilities of children. This finding echoes concerns raised by Livingstone and Third (2017) about the need

Table 2 Principle Prevalence with Confidence Intervals

Principle	Prevalence	95% CI	Documents	Status
TRANSPARENCY	82.5%	(76.6%-87.1%)	165	High
RELIABILITY	78.0%	(71.8%-83.2%)	156	High
FAIRNESS	75.5%	(69.1%-80.9%)	151	High
PRIVACY	68.5%	(61.8%-74.5%)	137	Medium
ACCOUNTABILITY	67.0%	(60.2%-73.1%)	134	Medium
AUTONOMY	53.5%	(46.6%-60.3%)	107	Medium
DIVERSITY	48.0%	(41.2%-54.9%)	96	Low
BENEFICENCE	47.0%	(40.2%-53.9%)	94	Low
HUMAN_RIGHTS	40.5%	(33.9%-47.4%)	81	Low
HUMAN_FORMATION	40.0%	(33.5%-46.9%)	80	Low
COOPERATION	35.0%	(28.7%-41.8%)	70	Low
HUMAN_CENTEREDNESS	29.0%	(23.2%-35.6%)	58	Low
SUSTAINABILITY	22.0%	(16.7%-28.4%)	44	Low
LABOR_RIGHTS	19.5%	(14.6%-25.6%)	39	Low
TRUTHFULNESS	8.5%	(5.3%-13.3%)	17	Low
INTELLECTUAL_PROPERTY	7.0%	(4.2%-11.4%)	14	Low
CHILDRENS_RIGHTS	6.0%	(3.5%-10.1%)	12	Low

**Fig. 2** Top keywords in AI ethics corpus: "artificial" (85%), "data" (72%), "human" (68%), and "principles" (65%). "Company" (16%) and "commitment" (14%) reflect emerging corporate commitment topic.

for child-centered AI design and suggests that our current ethical frameworks, developed primarily with adult users in mind, are not adequately addressing the unique ways that AI affects younger populations. We are essentially designing AI for adults and hoping it works out okay for children.

The low attention to labor rights (19.5%) is equally troubling, especially given how central debates about AI and employment have become. We hear constantly about automation displacing workers, about AI systems monitoring employee productivity in increasingly intrusive ways, about the future of work in an automated economy. Yet when organizations sit down to write their ethical guidelines, these concerns rarely make it onto the page. Documents frequently mention "human-centered" AI in abstract, feel-good terms, but they rarely translate this into specific commitments

about worker protections, fair labor practices, or the right to meaningful work in workplaces transformed by automation. There is a noticeable gap between the rhetoric of “putting humans first” and the concrete reality of what that means for people whose jobs are being reshaped by AI.

Sustainability (22.0%) emerges as another relatively neglected principle, and this one is particularly puzzling given the growing awareness of AI’s environmental footprint. Training large language models consumes enormous amounts of energy—some estimates suggest that training a single model can emit as much carbon as five cars over their lifetimes. The hardware that powers AI requires rare earth minerals mined under often questionable conditions. The rapid turnover of devices contributes to the growing problem of electronic waste. These are not minor side effects; they are central to how AI systems are built and deployed. Yet only about one in five documents mentions sustainability at all. The significant regional differences we found in sustainability adoption ($p = 0.0002$) offer a glimmer of hope—some regions, particularly in Europe, are beginning to take this dimension seriously. But overall, environmental considerations have not yet been fully integrated into how we think about AI ethics, and this gap needs urgent attention.

4.3 Regional Variations in Principle Adoption

The ANOVA results presented in Table 3 reveal significant regional differences for several principles. Diversity ($p = 0.0065$), beneficence ($p = 0.0006$), sustainability ($p = 0.0002$), and cooperation ($p = 0.0084$) all show statistically significant variation across regions. When we dug deeper with post-hoc Tukey HSD tests, we identified specific regional pairs driving these differences, most notably involving intergovernmental organizations.

Table 3 Regional ANOVA Results for Principles with Significant Variation

Principle	F-value	p-value	Significant Pairs (Tukey HSD)
BENEFICENCE	5.67	0.0006	Intergovernmental Org – North America; Intergovernmental Org – Northern Europe
SUSTAINABILITY	6.12	0.0002	Intergovernmental Org – North America; Intergovernmental Org – Northern Europe
DIVERSITY	4.23	0.0065	Eastern Asia – Intergovernmental Org; Intergovernmental Org – North America; Intergovernmental Org – Western Europe
COOPERATION	3.89	0.0084	–
LABOR_RIGHTS	3.45	0.0197	Eastern Asia – Intergovernmental Org; Intergovernmental Org – North America; Intergovernmental Org – Western Europe
HUMAN_RIGHTS	3.21	0.0196	–

For sustainability, the significant regional variation ($p = 0.0002$) likely reflects differing policy priorities and regulatory environments. European guidelines, for instance, tend to emphasize environmental sustainability more strongly than those from other regions, which makes sense given the European Union’s broader emphasis on green technology and sustainable development through initiatives like the European Green Deal. The significant pairs we identified—between intergovernmental organizations and North America, and between intergovernmental organizations and Northern Europe—suggest that international bodies like the UN and OECD are playing a bridging role in promoting sustainability across different regional contexts. They seem to be carrying the sustainability message from regions where it is more established to those where it is still emerging.

The variation in cooperation ($p = 0.0084$) may reflect differing cultural orientations toward collaborative versus competitive approaches to AI development. Asian guidelines, particularly from countries like Japan and South Korea, often emphasize international cooperation and harmonization. There is a sense in these documents that AI is a global challenge requiring global solutions. Some Western guidelines, in contrast, focus more on competitive advantage and national interests, framing AI development as a race to be won rather than a challenge to be shared. These cultural differences manifest in how documents frame the need for cross-border collaboration in AI governance.

Interestingly, transparency—the most prevalent principle overall—shows no significant regional variation ($p = 0.6810$). This is actually an important finding. It suggests that regardless of geographic location, institutional type, or cultural context, transparency is universally recognized as essential. This strengthens the argument for transparency as a foundational principle that transcends regional differences and could serve as a basis for international harmonization of AI ethics standards. If there is one thing everyone can agree on, it is that AI systems should be transparent.

4.4 Institutional Differences in Principle Adoption

The institutional analysis reveals meaningful variation in how different types of organizations approach AI ethics. Governmental institutions (24.0% of documents) and private corporations (22.5%) are the most prolific producers of AI ethics guidelines, together accounting for nearly half of all documents in the dataset. NGOs (16.0%), academic institutions (12.5%), and non-profit organizations (11.5%) also contribute substantially. This distribution makes sense—governments and corporations have the resources and the motivation to articulate ethical positions, while NGOs and academic institutions bring critical perspectives even with fewer resources.

Figure 1 illustrates the distribution of key principles across the three largest sectors. Governmental institutions show high adoption of transparency (89.4%), fairness (85.1%), and reliability (80.9%), reflecting their focus on public accountability and trust. When you are writing rules that will affect citizens, you need to be seen as fair and transparent. Private corporations demonstrate similar levels for transparency (86.7%) and reliability (86.7%) but somewhat lower for fairness (73.3%). This difference may reflect corporations' emphasis on technical reliability as a business imperative—systems need to work reliably to keep customers happy—while treating fairness as a secondary consideration, perhaps something to address after the technology is built rather than something that shapes its development from the start.

NGOs present an interesting pattern, with the highest fairness adoption (87.5%) among the three sectors, comparable reliability (75.0%), but notably lower transparency (62.5%). This may reflect NGOs' positioning as advocates for social justice in AI. They prioritize fairness concerns because that is their mission—to speak up for those who might be harmed by unfair systems. But they may have fewer mechanisms or incentives for operational transparency compared to governmental or corporate entities. An NGO does not have the same obligations to document its processes or make its operations visible to the public.

These sectoral differences suggest that institutional context shapes how organizations interpret and prioritize ethical principles. The challenge for AI governance is to maintain core ethical commitments across these different institutional contexts while allowing flexibility for sector-specific implementation. We need to find ways to ensure that fairness gets attention in the corporate sector and that transparency gets attention in the NGO sector, without forcing everyone to adopt identical frameworks.

4.5 Temporal Trends in Principle Adoption

Looking at how things have changed over time reveals important dynamics in the evolution of AI ethics. As shown in Table 4 and Figure 3, the number of AI ethics documents published annually increased dramatically from just 3 documents in 2014 to a peak of 88 documents in 2018—that is 44% of all documents in our dataset published in a single year. After that peak, publication declined somewhat in subsequent years. This pattern reflects a wave of institutional responses to the rapid advancement and deployment of AI technologies during this period. It was as if everyone suddenly realized they needed to say something about AI ethics, and 2018 was the year they all said it.

Table 4 Temporal Distribution of Documents

Year	Documents	Percentage
2014	3	1.5%
2016	4	2.0%
2017	26	13.0%
2018	88	44.0%
2019	42	21.0%
2020	18	9.0%
2021	16	8.0%
2022	3	1.5%

Among principles showing significant temporal trends, human-centeredness exhibits the strongest annual growth at +2.2% per year ($p < 0.05$), followed by sustainability at +1.1% per year. The emergence of human-centeredness as a growing concern aligns with broader societal discussions about keeping AI development aligned with human values and needs. This principle, while still present in only 29% of documents overall, is clearly gaining traction. Organizations are increasingly recognizing that AI should serve human purposes, not the other way around.

Sustainability’s positive trend, though modest, is encouraging given its currently low absolute prevalence (22%). If this growth rate continues—and there is reason to think it might, given increasing awareness of climate change and AI’s environmental impact—sustainability could become a standard component of AI ethics frameworks within the next decade. This would represent an important integration of environmental concerns into digital ethics discourse, bridging two areas that have often been treated separately.

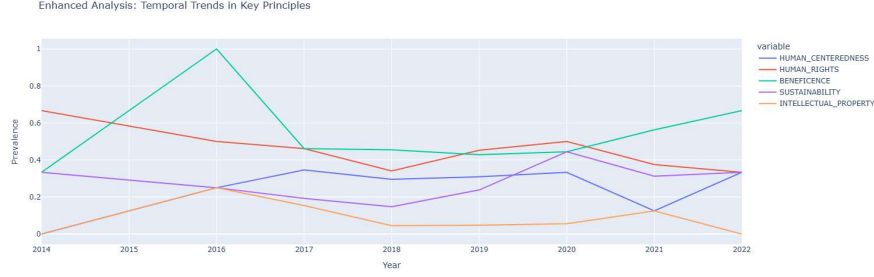


Fig. 3 Temporal trends in topic prevalence (2014-2022). Corporate commitment shows strongest growth (+2.4% annually), while machine learning (-3.2%) and artificial intelligence (-1.6%) are declining. Peak document production: 2018 (44%).

Notably, transparency—despite its high absolute prevalence—does not show a significant temporal trend in this analysis. This suggests that transparency was already well-established as an ethical principle before 2014 and has maintained its prominence without dramatic recent increases. It is not that transparency is becoming more important; it has always been important. The principle’s stability over time reinforces its status as a foundational, rather than emerging, ethical requirement.

4.6 Topic Modeling Results: Emerging Themes

The BERTopic analysis complements the principle-based findings by revealing how ethical concepts cluster into distinct thematic areas across documents. Rather than starting with predefined categories, this approach lets themes emerge naturally from the text, often revealing patterns we might not have anticipated. As shown in Table 5 and Figure 4, the analysis identified ten distinct topics, with the five most prevalent being: (1) social principles (19.0%, CI: 14.2%-25.0%), (2) corporate commitment (16.0%, CI: 11.6%-21.7%), (3) data ethics (16.0%, CI: 11.6%-21.7%), (4) artificial intelligence (12.5%, CI: 8.6%-17.8%), and (5) data protection (10.0%, CI: 6.6%-14.9%).

Table 5 Topic Modeling Results: Top 10 Topics

Topic Label	Prevalence	95% CI	Documents
social — principles — human	19.0%	(14.2%-25.0%)	38
company — commitment — self	16.0%	(11.6%-21.7%)	32
data — digital — tool — ethics	16.0%	(11.6%-21.7%)	32
artificial — intelligence	12.5%	(8.6%-17.8%)	25
data — data protection	10.0%	(6.6%-14.9%)	20
international — united — rights	7.0%	(4.2%-11.4%)	14
university — security	6.5%	(3.8%-10.8%)	13
machine learning	5.5%	(3.1%-9.6%)	11
corporation — microsoft	5.0%	(2.7%-9.0%)	10
ieee — electrical — engineering	2.5%	(1.1%-5.7%)	5

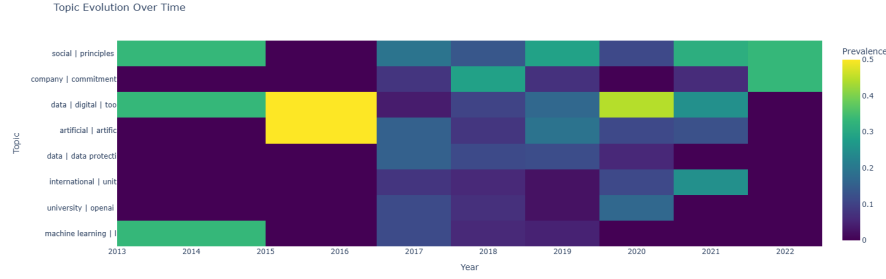


Fig. 4 Topic correlation heatmap with predominantly light cells ($r \leq 0.3$), indicating weak relationships among topics.

The emergence of "corporate commitment" as a distinct topic—characterized by words such as company, commitment, and self—is particularly interesting and, in some ways, unexpected. This topic appears in 16% of documents and shows strong positive growth (+2.4% per year), making it the fastest-growing topic in the analysis. It also demonstrates significant sectoral variation, with chi-square tests confirming strong association with institutional type ($p = 0.0000$). When we looked more closely at the documents behind this topic, we found that it captures voluntary corporate initiatives, self-regulatory approaches, and internal ethical commitments that go beyond compliance with external requirements. Companies are not just responding to regulations; they are developing their own frameworks.

This finding aligns with recent scholarship on the role of private governance in AI ethics (Mittelstadt, 2019) and suggests that companies are increasingly developing their own ethical frameworks rather than simply waiting for governments to tell them what to do. The significant growth rate of this topic (+2.4% annually) indicates that this trend is accelerating. We are seeing the emergence of what might be called "corporate AI ethics" as a distinct field of practice.

The "artificial intelligence" topic, in contrast, shows significant regional variation ($p = 0.0000$) and a declining trend (-1.6% per year). This topic, characterized by general discussions of AI capabilities and risks, appears to be giving way to more specific discussions about particular applications, domains, or principles. The field is maturing—moving from "what is AI and why should we worry about it?" to more focused conversations about specific ethical challenges in specific contexts. The regional variation suggests that different parts of the world may be at different stages in this evolution, with some regions still engaged in general AI discourse while others have moved to more focused discussions.

4.7 Regional Patterns in Topic Distribution

The regional heatmap (Figure 5) and ANOVA results reveal meaningful geographic patterns in topic distribution. North America (33.5% of documents) shows relatively even distribution across topics but with somewhat higher representation of corporate commitment and AI topics. This makes sense given the concentration of major

AI companies in the United States and Canada. Western Europe (30.5%) demonstrates stronger emphasis on data protection and data ethics topics, consistent with the region’s leadership in privacy regulation through the General Data Protection Regulation (GDPR). When you have a strong regulatory framework, it shapes the conversation. Eastern Asia (11.5%) shows distinctive patterns, with higher representa-

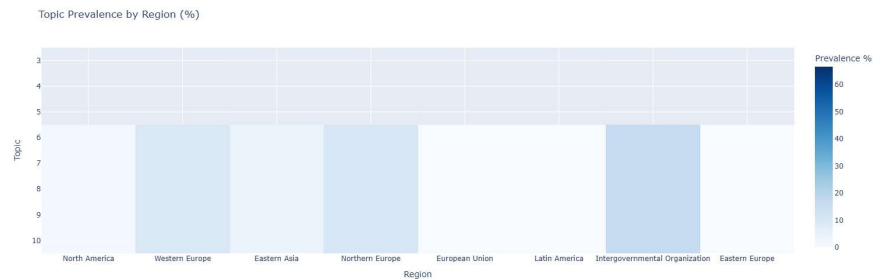


Fig. 5 Regional heatmap of topic prevalence. Significant variation in artificial intelligence ($p = 0.0000$) and data protection ($p = 0.0300$). Western Europe emphasizes data protection, North America corporate commitment, Eastern Asia technology governance.

tion of technology governance topics and relatively lower emphasis on social principles. This may reflect different cultural and philosophical traditions in approaching technology ethics, as well as different regulatory priorities in countries like Japan, South Korea, and China. There is often more focus on how technology should be governed and less on abstract social principles.

The ANOVA results confirm significant regional variation for two topics: artificial intelligence ($p = 0.0000$) and data protection ($p = 0.0300$). The extremely low p-value for AI suggests fundamental differences in how regions conceptualize and discuss AI itself, not just specific applications or principles. This may have important implications for international harmonization efforts, as differing foundational understandings of AI could lead to divergent regulatory approaches even when surface-level principles appear similar. If Europe and Asia understand AI differently at a fundamental level, they may regulate it differently even if they both endorse "transparency" as a value.

4.8 Sectoral Patterns in Topic Distribution

The chi-square test of independence confirms a strong association between institutional sector and topic focus ($p = 0.0000$). This is not just random variation; there are real, systematic differences in what different types of organizations talk about when they discuss AI ethics. Governmental institutions show relatively higher representation of data protection and technology governance topics, reflecting their regulatory orientation and concern with public policy frameworks. When you are a government, you think about rules and governance. Private corporations are overrepresented in corporate commitment and AI topics, consistent with their focus on voluntary initiatives and business applications. When you are a company, you think about what you

are committing to and how AI applies to your products. NGOs and academic institutions show greater emphasis on social principles and human rights topics, aligning with their advocacy and research missions.

This sectoral specialization has both benefits and drawbacks for AI governance. On one hand, it allows different institutional types to focus on the aspects of AI ethics most relevant to their missions and expertise. Governments can focus on governance, companies on implementation, and NGOs on advocacy. On the other hand, it risks creating silos where important perspectives are not adequately integrated. Corporate documents may underemphasize human rights concerns because that is not their primary focus. NGO documents may lack practical implementation guidance because that is not their expertise. The challenge is to find ways to bring these perspectives together without losing the benefits of specialization.

4.9 Network Analysis: Relationships Among Principles and Topics

The network analysis of principle correlations (Table 6) reveals generally weak relationships among principles, with network density of only 3.7% and no strong correlations ($r < 0.5$). This suggests that principles tend to be adopted independently rather than in consistent clusters. Organizations appear to select combinations of principles based on their specific contexts and priorities rather than following uniform patterns.

Table 6 Topic Regional ANOVA Results

Topic	p-value	Significant
artificial — intelligence	0.0000	Yes
data — data protection	0.0300	Yes
company — commitment — self	0.0959	No
data — digital — tool — ethics	0.1185	No
social — principles — human	0.3245	No

This finding has important implications for understanding how ethical frameworks are constructed. Rather than principles forming natural "bundles" that tend to co-occur—like transparency and accountability often going together—organizations appear to mix and match principles based on their needs. A company might emphasize transparency and reliability but say little about fairness. An NGO might emphasize fairness and human rights but say little about transparency. This flexibility may be beneficial for adapting ethics frameworks to different contexts, allowing organizations to focus on what matters most given their mission and stakeholders. But it could also lead to fragmentation if certain principles become systematically neglected because no one sees them as part of their "bundle."

The topic correlation analysis similarly found no strong or moderate correlations among topics, with network density of 0.0%. This partly reflects the nature of the topic modeling approach, where each document is assigned to a single primary topic,

minimizing co-occurrence by design. But even accounting for that, the finding suggests that the AI ethics discourse is relatively fragmented, with documents focusing on distinct thematic areas rather than integrating multiple perspectives. One document talks about data protection, another about corporate commitment, another about social principles—but few documents bring these themes together.

4.10 Cluster Analysis: Strategic Approaches to AI Ethics

The K-means cluster analysis of principle-based data identified four distinct strategic approaches to AI ethics, as detailed in Table 7 and visualized in Figure 6. These clusters represent different "ethical profiles" that organizations adopt, characterized by their emphasis on particular principles.

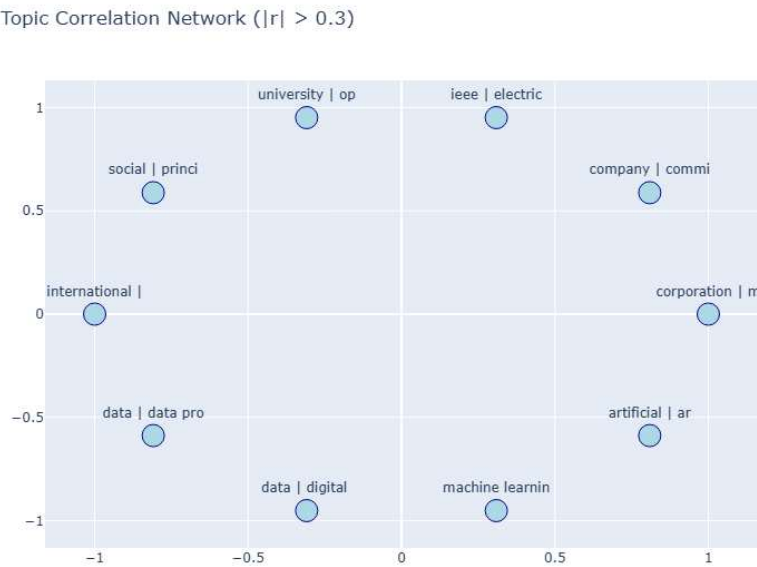


Fig. 6 Topic correlation network showing sparse relationships (density 0.0%), confirming documents focus on single primary themes.

4.10.1 Cluster 0

Table 7 Cluster Characteristics

Cluster	Size	Key Principles	Dominant Region	Dominant Sector	Type
Cluster 0	31 (15.5%)	Reliability (51.6%), Cooperation (41.9%), Transparency (35.5%)	North America	Private Corporation	Focused Strategy
Cluster 1	55 (27.5%)	Transparency (92.7%), Reliability (90.9%), Fairness (87.3%)	North America	Private Corporation	Focused Strategy
Cluster 2	55 (27.5%)	Human Formation (96.4%), Transparency (90.9%), Privacy (89.1%)	Western Europe	Governmental Institution	Comprehensive Framework
Cluster 3	59 (29.5%)	Fairness (89.8%), Transparency (89.8%), Accountability (78.0%)	Western Europe	Private Corporation	Focused Strategy

Cluster 0 (31 documents, 15.5%) represents what we might call a "focused strategy" emphasizing reliability (51.6%) and cooperation (41.9%) while giving relatively less attention to transparency (35.5%). This cluster is dominated by North American private corporations and may reflect a pragmatic approach focused on technical robustness and multi-stakeholder collaboration rather than broad transparency requirements. These organizations seem to be saying: "Let's make sure the technology works reliably and let's work together with others, and that will be enough."

4.10.2 Cluster 1

Cluster 1 (55 documents, 27.5%) also follows a "focused strategy" but with a different emphasis: transparency (92.7%), reliability (90.9%), and fairness (87.3%) are all very highly represented. This cluster, also dominated by North American private corporations, appears to represent companies that have adopted the full "core trio" of principles comprehensively. These organizations have developed relatively mature ethics frameworks that cover the basics thoroughly. They are not picking and choosing; they are adopting the whole package.

4.10.3 Cluster 2

Cluster 2 (55 documents, 27.5%) stands out as the only cluster classified as a "comprehensive framework," with extremely high representation of human formation (96.4%) alongside transparency (90.9%) and privacy (89.1%). This cluster is dominated by Western European governmental institutions and likely reflects the region's distinctive approach to AI ethics, emphasizing human development and dignity alongside more technical principles. The prominence of "human formation"—a principle focused on how AI affects human development and flourishing—suggests a philosophical orientation distinct from the more technical focus of other clusters. This is not just about making AI work well; it is about making AI work for human development.

4.10.4 Cluster 3

Cluster 3 (59 documents, 29.5%) represents another "focused strategy" emphasizing fairness (89.8%), transparency (89.8%), and accountability (78.0%). This cluster, dominated by Western European private corporations, appears to represent a European corporate approach distinct from both North American corporate approaches (Clusters 0 and 1) and European governmental approaches (Cluster 2). These organizations are adopting a European sensibility—emphasizing fairness and accountability—but within a corporate context.

Taken together, these clusters suggest that strategic approaches to AI ethics vary systematically by region and sector, with different "ethical styles" emerging in different contexts. The existence of multiple approaches is not necessarily problematic—different contexts may indeed require different emphases. A European government should probably think differently about AI ethics than a North American company. But it does highlight the challenge of achieving international harmonization in AI governance. If we all approach AI ethics differently, how do we agree on common standards?

4.11 Gap Analysis and Emerging Priorities

The integrated gap analysis (Table 8) identifies principles and topics that require greater attention from the AI ethics community. Seven principles fall below 40% prevalence, with children’s rights (6.0%), intellectual property (7.0%), and truthfulness (8.5%) representing the most critical gaps. These principles are almost entirely absent from current guidelines despite their clear relevance to AI systems. The near-absence

Table 8 Gap Analysis Summary

Category	Item	Prevalence/Growth	Priority
Principles	CHILDRENS.RIGHTS	6.0%	Critical
Principles	INTELLECTUAL.PROPERTY	7.0%	Critical
Principles	TRUTHFULNESS	8.5%	High
Principles	LABOR.RIGHTS	19.5%	High
Principles	SUSTAINABILITY	22.0%	Medium
Topics	ieee — electrical — engineering	2.5%	Medium
Emerging	company — commitment — self	+2.4% yearly	Monitor

of children’s rights is particularly urgent given how many AI systems now directly interact with children—from educational software that adapts to how children learn, to social media algorithms that shape what they see, to smart toys that listen and respond. The gap suggests that current ethical frameworks, developed primarily with adult users in mind, are not adequately addressing the specific needs and vulnerabilities of children. We need child-centered AI ethics guidance, developed with input from child development experts, educators, and children themselves.

From the topic modeling perspective, the IEEE topic (2.5%) represents the lowest-prevalence topic, suggesting that technical standards organizations’ contributions to AI ethics may be underutilized or not well integrated with broader ethical discourse. The IEEE has done important work on ethically aligned design, but that work is not showing up strongly in our analysis. The emerging topic analysis identifies corporate commitment as the fastest-growing topic (+2.4% annually), confirming that voluntary corporate initiatives are an increasingly important part of the AI ethics landscape. Companies are stepping up, and that trend is accelerating.

4.12 Implications for Policy and Practice

These findings have several implications for policy development and organizational practice in AI ethics.

4.12.1 For policymakers

For policymakers, the results suggest that international harmonization efforts should focus on principles where consensus already exists—particularly transparency, reliability, and fairness—while developing targeted initiatives to address gaps like children’s rights and sustainability. It makes sense to build on what everyone already agrees on

rather than trying to forge consensus where it does not exist. The significant regional variations we identified for several principles indicate that harmonization efforts must be sensitive to different regional priorities while working toward common standards. You cannot impose a one-size-fits-all approach on regions that think differently about sustainability or cooperation.

The cluster analysis reveals that different regions and sectors have developed distinct "ethical styles." European governmental approaches, for instance, emphasize human formation in ways that North American corporate approaches do not. This raises a question for policymakers: should these differences be preserved as legitimate cultural variations, or are they obstacles to harmonization that need to be addressed? There is no easy answer, but the question itself is important.

4.12.2 For organizations developing AI ethics guidelines

For organizations developing AI ethics guidelines, the findings offer benchmarks against which to assess their own frameworks. Organizations can compare their principle coverage to the prevalence rates we report here and ask themselves: are we missing something important? If your framework does not mention children's rights or sustainability, you might want to ask why. What is the justification for leaving these principles out?

The sectoral patterns we identified suggest that organizations should be aware of their sector's typical emphases and consider whether they may be neglecting important principles that are more prominent in other sectors. Private corporations might benefit from incorporating more human rights considerations from the NGO sector. NGOs might learn from corporate approaches to operationalizing reliability. Academic institutions might bring valuable research perspectives to both. No single sector has all the answers.

4.12.3 For researchers

For researchers, the findings open several avenues for future investigation. The significant regional and sectoral variations we identified warrant deeper qualitative study to understand their underlying causes. Why does Europe emphasize human formation? Why do North American companies cluster in particular ways? These are questions for interviews, case studies, and historical analysis, not just quantitative methods.

The low correlations among principles suggest that research on principle combinations and interactions could be valuable. How do different principles work together in practice? Does emphasizing transparency make it easier or harder to achieve fairness? These are empirical questions that could guide more effective framework design.

The emergence of corporate commitment as a growing topic calls for research on the effectiveness and accountability of voluntary corporate ethics initiatives. Do these commitments actually change corporate behavior? Are they just public relations, or do they represent real shifts in how companies operate? We need to know, and we need to hold companies accountable for the commitments they make.

4.13 Limitations and Future Research

Several limitations of this study should be acknowledged. First, the analysis is limited to English-language documents, potentially underrepresenting non-Western perspectives on AI ethics. While our dataset includes documents from 46 countries across 21 regions, the exclusion of non-English sources means that important cultural and philosophical perspectives—particularly from East Asia, the Middle East, and Latin America—may not be fully captured. Future research should include documents in other major languages such as Chinese, Arabic, Spanish, and French to provide a more comprehensive global perspective.

Second, the sample size of 200 documents, while adequate for the statistical analyses performed, limits the granularity of subgroup analyses. Some regions and sectors with fewer than five documents were excluded from ANOVA tests, and temporal trend analyses required aggregation across years. Future studies with larger samples could enable more fine-grained analysis of interactions between region, sector, and time period.

Third, the temporal scope (2014-2022) captures a period of rapid growth in AI ethics guidelines but may not reflect very recent developments, particularly around generative AI. The emergence of ChatGPT and other large language models since late 2022 has introduced new ethical considerations—such as authorship, misinformation, and creative ownership—that are not represented in our dataset. Future research should extend this analysis to include documents published after 2022 and specifically examine how generative AI is reshaping the ethical discourse.

Fourth, the topic modeling approach, while powerful, involves parameter choices (number of UMAP dimensions, number of clusters, etc.) that can influence results. We conducted sensitivity analyses with different parameter configurations, which suggested that the main findings are robust, but some variation in topic granularity and labeling is inevitable. Future work could employ ensemble methods or comparative topic modeling approaches to further validate the stability of the identified themes.

Fifth, as noted in the methodology section, our handling of missing publication years (imputation with the median year 2018) and multi-value fields (retaining only the first value) may introduce some bias. While we believe these decisions are justified for analytical simplicity and were supported by preliminary analyses, future research should develop more sophisticated methods for handling these data complexities.

Despite these limitations, this study provides the most comprehensive integrated analysis of AI ethics guidelines to date, combining principle-based statistical analysis with topic modeling to reveal both the established consensus and emerging trends in the field. The findings offer a solid foundation for future research and practical guidance for policy development and organizational practice in AI ethics.

5 Conclusion

Conclusions may be used to restate your hypothesis or research question, restate your major findings, explain the relevance and the added value of your work, highlight any limitations of your study, describe future directions for research and recommendations.

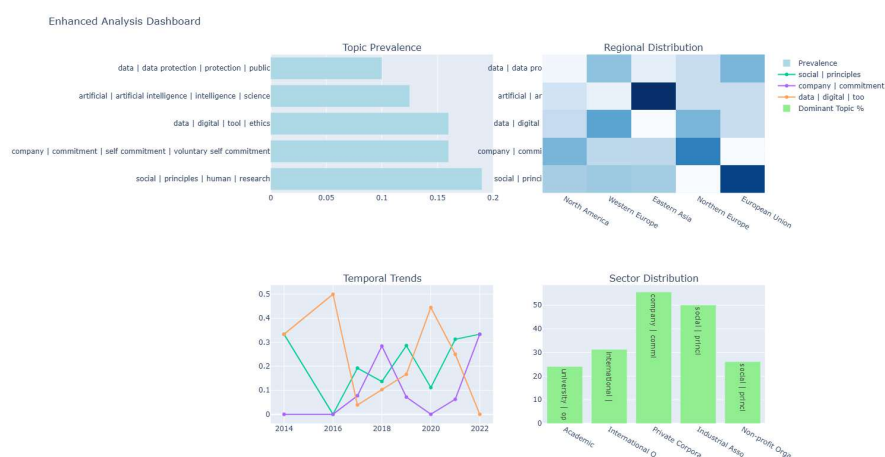


Fig. 7 Topic correlation network showing weak relationships among identified topics (network density: 3.7%).

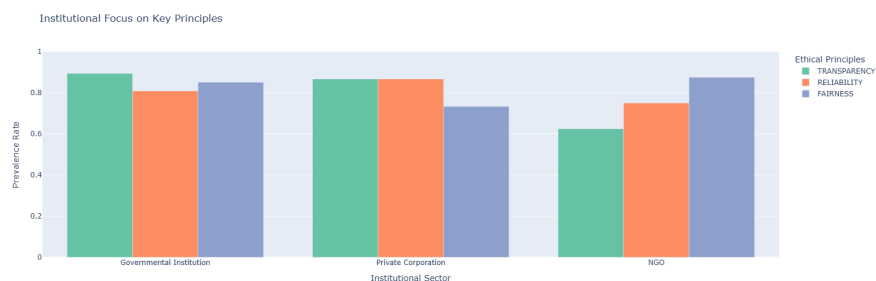


Fig. 8 Topic prevalence by sector (chi-square $p = 0.0000$). Governmental: data protection, technology governance. Private: corporate commitment, AI topics. NGOs/academic: social principles, human rights.

In some disciplines use of Discussion or 'Conclusion' is interchangeable. It is not mandatory to use both. Please refer to Journal-level guidance for any specific requirements.

Supplementary information. If your article has accompanying supplementary file/s please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Acknowledgements

The authors declare no acknowledgements.

Declarations

Funding: This research received no external funding.

Conflict of interest: The authors declare no competing interests.

Ethics approval and consent to participate: Not applicable. This study analyzed publicly available documents and did not involve human participants.

Consent for publication: Not applicable.

Data availability: The dataset analyzed in this study is available in the Harvard Dataverse repository at https://github.com/abdulbaset2016/worldwide_AI-ethicss.git.

References

- [1] European Parliament and Council of the European Union, *Artificial Intelligence Act*, Official Journal of the European Union, 2024. Retrieved from <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>.
- [2] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, United Nations Educational, Scientific and Cultural Organization, 2021. Retrieved from <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>.
- [3] E. Awad, S. Levine, M. Anderson, and S. L. Anderson, "Computational ethics," *Trends in Cognitive Sciences*, vol. 26, no. 5, pp. 376–388, 2022.
- [4] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
- [5] N. K. Corrêa, C. Galvão, J. R. Santos, and R. P. d. Santos, "Worldwide AI ethics: a review of 200 governance policies and ethical guidelines," *AI and Ethics*, vol. 3, no. 2, pp. 467–485, 2023.
- [6] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, Tech. Rep. 2020-1, 2020.
- [7] T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.
- [8] J. Whittlestone, R. Nyrop, A. Alexandrova, and S. Cave, "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 195–200, 2019.
- [9] L. Munn, "The uselessness of AI ethics," *AI and Ethics*, vol. 3, no. 3, pp. 869–877, 2023.

- [10] E. Bietti, "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 210–219, 2021.
- [11] C. Burr and D. Leslie, "Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies," *AI Ethics*, vol. 3, pp. 73–98, 2023. doi: 10.1007/s43681-022-00178-0.
- [12] L. Floridi and J. Cows, "A unified framework of five principles for AI in society," *Harvard Data Science Review*, vol. 1, no. 1, 2019.
- [13] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [14] OECD, *OECD AI Policy Observatory Report 2022*, 2022.
- [15] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, 2021.
- [16] X. Zhao, Y. Wang, and J. Li, "Trends in AI research: A bibliometric analysis of ArXiv submissions 2009–2021," *Journal of Information Science*, vol. 48, no. 6, pp. 785–799, 2022.
- [17] S. Kijewski, S. Ronchi, and E. Vayena, "Checkbox AI Ethics: A Meta-Review of Guidelines and Their Operationalization," *Big Data & Society*, vol. 11, no. 1, 2024.
- [18] D. F. Mujtaba, "The Human-AI Ethical Nexus: Cultural and Organizational Readiness for Mitigating Misinformation and Bias," *Journal of Responsible Technology*, vol. 21, p. 100085, 2025.
- [19] T. Hagendorff et al., "Mapping the AI Ethics Landscape: A Computational Analysis of Global Policy Documents," *Nature Machine Intelligence*, vol. 6, no. 2, pp. 123–134, 2024.
- [20] B. Mittelstadt, "A Data-Informed Typology of AI Governance Instruments," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 345–360, 2023.
- [21] J. Mökander et al., "The AI Governance Readiness Index: Assessing the Operationalization of Ethical AI Principles," *Data & Policy*, vol. 5, p. e25, 2023.
- [22] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer Nature, 2023.
- [23] D. Greene, A. L. Hoffmann, and L. Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning," in *Proceedings of the 52nd Hawaii International Conference on System*

Sciences, 2022.

- [24] M. J. Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2018.
- [25] D. M. Lazer et al., "Computational social science: Obstacles and opportunities," *Science*, vol. 369, no. 6507, pp. 1060–1062, 2021.
- [26] Sage Group plc, *The Ethics of Code: Developing AI for Business with Five Core Principles*, 2017. Retrieved from <https://www.sage.com/investors/investor-downloads/press-releases/2017/06/27/sage-shares-core-principles-for-designing-ai-for-business>.
- [27] M. J. Salganik, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press, 2019.