UNESCO AI Ethics Recommendation

Global Alignment Assessment Across 2,100+ Policies

Lucas Sempé

February 11, 2026

Table of contents

1 UNESCO AI Ethics Recommendation: Global Alignment Assessment

Measuring Policy Convergence with UNESCO Principles Across 2,100+ Documents

Preface

This book assesses global alignment with the UNESCO Recommendation on the Ethics of Artificial Intelligence — the first-ever global standard on AI ethics adopted by 193 Member States in November 2021.

Drawing on 2,100+ policies across 70+ jurisdictions, we measure how closely national policies align with UNESCO's ten core values and eleven action areas. The analysis reveals patterns of adoption, adaptation, and divergence from the UNESCO framework.

1.1 Key Findings

- **Moderate alignment**: Mean UNESCO score 1.68/4.0, indicating partial adoption
- **Value priorities**: Human rights and transparency emphasized; sustainability lagging
- **Cluster structure**: Policies divide into "Comprehensive Alignment" vs "Selective Adoption"
- **Temporal dynamics**: Post-2021 policies show stronger UNESCO alignment

1.2 Methodology

We employ an LLM ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) achieving ICC = 0.827 (excellent inter-rater reliability), scoring policies against UNESCO's 21-component framework.

Citation: Sempé, L. (2026). *UNESCO AI Ethics Recommendation: Global Alignment Assessment.* International Initiative for Impact Evaluation (3ie).

Data and Code: github.com/lsempe77/ai-governance-capacity

2 Introduction

2.1 Global Alignment with UNESCO AI Ethics

Chapter summary. This chapter introduces the UNESCO Recommendation on the Ethics of Artificial Intelligence and our measurement of global policy alignment with its framework. We examine adoption patterns, adaptation strategies, and divergence from the UNESCO standard.

2.1.1 The UNESCO Milestone

In November 2021, UNESCO's 193 Member States adopted the **Recommendation on the Ethics of Artificial Intelligence**—the first global normative instrument on AI ethics. The Recommendation establishes:

- **10 core values**: Human rights, environment, diversity, inclusiveness, peaceful societies, human oversight, transparency, responsibility, accountability, privacy
- **11 action areas**: Ethical impact assessment, governance, data policy, development cooperation, environment, gender, culture, education, communication, economy, regulation

This provides the most comprehensive multilateral framework for AI ethics governance, grounding national policies in shared principles while allowing contextual adaptation.

2.1.2 Research Questions

This book addresses three questions:

1. **How aligned** are national AI policies with UNESCO's framework?
2. **Which values and action areas** receive priority versus neglect?
3. **How has alignment evolved** since the Recommendation's adoption?

2.1.3 Measurement Framework

We score each policy on 21 UNESCO components (10 values + 11 action areas) using the same LLM ensemble methodology achieving ICC = 0.827. Scores reflect depth of engagement: 0 = absent, 1 = mentioned, 2 = described, 3 = operationalized, 4 = comprehensive.

2.1.4 Key Findings Preview

- **Moderate alignment**: Mean UNESCO score 1.68/4.0 (between "mentioned" and "described")
- **Value priorities**: Human rights (1.92) and transparency (1.85) lead; environmental sustainability (1.28) lags
- **Two-cluster structure**: "Comprehensive Alignment" (28%) vs "Selective Adoption" (72%)
- **Post-2021 increase**: Policies adopted after UNESCO Recommendation show stronger alignment

2.1.5 Roadmap

- **Chapter 17** maps UNESCO alignment landscape
- **Chapter 18** examines determinants of alignment
- **Chapter 19** identifies policy clusters and typologies
- **Chapter 20** traces temporal dynamics
- **Chapter 14** presents robustness checks
- **Chapters 15-16** discuss implications and conclusions

The analysis reveals selective rather than wholesale UNESCO adoption, with countries prioritizing values matching existing governance priorities.

3 Literature Review

3.1 Theoretical Foundations

> **Chapter summary.** We situate UNESCO alignment measurement within international norm diffusion literature, examining how global standards influence national policies and the mechanisms of regulatory convergence versus divergence.

3.1.1 The UNESCO Recommendation

In November 2021, UNESCO's 193 Member States adopted the **Recommendation on the Ethics of Artificial Intelligence**—the first global normative instrument on AI ethics. The Recommendation establishes:

- **10 core values**: Human rights, environment, diversity, inclusiveness, peaceful societies, human oversight, transparency, responsibility, accountability, privacy
- **11 action areas**: Ethical impact assessment, governance, data policy, development cooperation, environment, gender, culture, education, communication, economy, regulation

(**taddeo2021?**) characterizes the UNESCO Recommendation as representing "soft law"—non-binding but influential through moral authority and coordination effects. (**floridi2021?**) argues it provides the most comprehensive multilateral AI ethics framework, integrating principles from diverse traditions.

3.1.2 International Norm Diffusion

3.1.2.1 Mechanisms of Regulatory Convergence

(**simmons2006?**) identifies four mechanisms driving policy convergence:

1. **Coercion**: Powerful actors impose rules (World Bank conditionality)
2. **Competition**: Regulatory arbitrage pressures jurisdictions toward common standards
3. **Learning**: Countries emulate successful policies from peers
4. **Emulation**: Mimetic isomorphism driven by legitimacy rather than function

Our analysis examines which mechanisms drive UNESCO alignment. If coercion or competition dominates, we expect wealth-driven adoption patterns. If learning or emulation dominates, we expect horizontal diffusion within regions or income groups.

3.1.2.2 The Brussels Effect

Bradford (2020) theorizes the "Brussels Effect"—EU regulations becoming global standards through market power. In data protection, GDPR created de facto global norms as firms adopted EU standards globally rather than maintaining separate compliance systems.

But AI ethics differs from data protection: no single jurisdiction dominates global AI markets sufficiently to impose standards unilaterally. UNESCO's multilateral framework thus operates through different diffusion mechanisms than GDPR, emphasizing learning and legitimacy over market coercion.

3.1.3 Global Standards and Local Adaptation

(**acharya2004?**) distinguishes **localization** (modifying global norms to fit local contexts) from **transplantation** (wholesale adoption). (**wiener2020?**) shows that international norms rarely transplant unchanged—they undergo reinterpretation reflecting local values, institutions, and priorities.

Our measurement framework captures this by scoring **depth of engagement** rather than binary adoption: policies can mention UNESCO values (score 1), describe them contextually (score 2), operationalize them through requirements (score 3), or establish comprehensive governance (score 4). This allows distinguishing superficial from substantive adoption.

3.1.4 Value Priorities and Selectivity

Jobin, Ienca, and Vayena (2019) documents convergence on core AI ethics principles but divergence on prioritization. (**winfield2021?**) shows regional variation: European frameworks emphasize human rights and fundamental freedoms, Asian frameworks emphasize social harmony and collective welfare, Middle Eastern frameworks emphasize cultural values and religious principles.

UNESCO's framework accommodates this diversity through breadth—including 10 values and 11 action areas—allowing countries to prioritize different components while claiming UNESCO alignment. Our analysis examines whether countries adopt the full framework or selectively emphasize values matching existing priorities.

3.1.5 Implementation Challenges

(**cihon2021?**) documents challenges translating UNESCO principles into national governance:

- **Vagueness**: UNESCO values stated broadly, requiring national specification
- **Comprehensiveness**: 21 components (10 values + 11 action areas) create implementation burden
- **Novelty**: Action areas like "ethical impact assessment" lack established implementation templates
- **Coordination**: Multiple government agencies must implement different action areas

These challenges predict variable UNESCO alignment even among committed member states. Our measurement distinguishes rhetoric (mentioning UNESCO) from implementation (operationalizing UNESCO components).

3.1.6 Research Questions

This literature motivates three questions:

1. **How aligned** are national policies with UNESCO's 21-component framework?
2. **Which values and action areas** receive priority versus neglect?
3. **Has alignment increased** since the Recommendation's 2021 adoption?

3.1.7 Contribution

This book provides:

1. **Comprehensive measurement**: First systematic scoring of UNESCO alignment across 2,100+ policies
2. **Component-level analysis**: Distinguishing which UNESCO elements nations adopt versus ignore
3. **Temporal assessment**: Testing whether post-2021 policies show stronger alignment
4. **Diffusion patterns**: Examining whether UNESCO spreads through learning, competition, or legitimacy

The following chapters reveal selective rather than comprehensive UNESCO adoption, with countries prioritizing values matching existing governance frameworks.

4 Data & Methods

4.1 The OECD.AI Corpus

Chapter summary. This chapter describes the data collection pipeline: from the OECD.AI Policy Observatory through document retrieval, text extraction, and quality classification. We detail the construction of a 2,216-policy corpus with 11.4 million words of analysis-ready text across 70+ jurisdictions.

4.1.1 Data Source

Our data come from the **OECD.AI Policy Observatory** (OECD 2024), the most comprehensive international tracker of AI policy initiatives. Established as a collaborative effort among OECD member states and partner countries, the Observatory serves as the global standard for monitoring AI governance activity. It catalogues government actions related to AI — including national strategies, legislation, executive orders, guidelines, and programmes — with structured metadata on jurisdiction, year, policy type, target sectors, and responsible organisations. This structured approach makes the Observatory uniquely suited for systematic cross-national comparison, as each entry follows a consistent documentation schema that enables quantitative analysis at scale.

We politely scraped the complete Observatory as of January 2026, obtaining **2,216 policy entries** spanning **70+ jurisdictions** and the years **2017–2025**. This snapshot represents the state of global AI governance at a critical juncture, as many jurisdictions transition from voluntary guidelines to binding regulation.

Table 4.1: Corpus overview

Metric	Value
Total policy entries	2,216
Unique jurisdictions	70+
Time span	2017–2025
Policy types	Strategies, laws, guidelines, executive orders, programmes
Source	OECD.AI Policy Observatory

Table 4.1 shows the breadth of our corpus, which encompasses nearly every documented AI governance initiative globally over the past eight years. The 70+ jurisdictions include not only major

economies but also developing countries in Africa, Asia, and Latin America, providing the geographic diversity necessary to examine capacity gaps across income levels.

4.1.2 Document Retrieval

The OECD.AI Observatory provides brief descriptions (typically <500 words) and links to source documents, but does not host full texts. This design reflects the Observatory's role as a catalog rather than an archive — it points to official documents but leaves them at their original locations. For our analysis, however, we required the complete policy texts to enable detailed assessment of implementation capacity. This necessitated building a retrieval pipeline capable of locating and downloading documents that might have moved, been renamed, or disappeared from their original URLs.

Our five-strategy retrieval pipeline operated as a cascading fallback system. First, we attempted direct downloads from the `source_url` field provided in the Observatory metadata, which succeeded for approximately 60% of entries. For documents where direct download failed, we scraped the OECD.AI web page for each policy entry to locate embedded source links that might not appear in the structured metadata. When original URLs had moved or expired — a common occurrence for policy documents published years earlier — we queried the Internet Archive Wayback Machine to retrieve historical snapshots. For documents unavailable through any of these channels, we conducted targeted searches using DuckDuckGo with carefully constructed queries combining the policy title, jurisdiction, and file type restrictions. Finally, for the most difficult cases, we employed the Claude API's web search capability to locate official document URLs through more sophisticated reasoning about likely hosting locations.

This layered approach achieved approximately 94% coverage, successfully retrieving around 2,085 documents to local storage. The remaining entries — primarily press releases, brief announcements, or policies documented only through secondary sources — remained available as OECD snippets, providing at least minimal text for analysis even when full documents proved inaccessible.

4.1.3 Text Extraction

Retrieving documents was only the first challenge; extracting clean, analysis-ready text from diverse file formats proved equally demanding. Policy documents arrive in varied formats — PDFs may be text-based or scanned images, web pages may embed content within complex navigation structures, and documents may span from single-page executive summaries to hundred-page legislative texts. Each format required specialized handling to extract content accurately while removing headers, footers, page numbers, and other non-substantive elements that would interfere with analysis.

We developed format-specific extraction pipelines matched to document characteristics. For PDF documents — the most common format in our corpus — we employed PyMuPDF (`fitz`), which excels at extracting text from text-based PDFs while preserving document structure. For HTML documents, we used `trafilatura`, a content extraction library specifically designed to identify main textual content while stripping navigation menus, sidebars, and other boilerplate elements typical of government websites. For entries where no downloadable source could be located, we fell back

to the OECD snippet text, accepting the limitation of abbreviated content rather than excluding these policies entirely.

Each document was then classified into one of three quality tiers based on extracted word count, providing a systematic approach to assessing text adequacy for detailed analysis:

Table 4.2: Text quality distribution

Quality Tier	Word Count	N	%	Description
Good	500 words	948	42.8%	Full analysis possible
Thin	100–499 words	806	36.4%	Usable with caveats
Stub	<100 words	462	20.8%	Minimal text only
	Analysis-ready	**1,754**	**79.2%**	Good + Thin

Table 4.2 reveals that nearly 80% of our corpus (1,754 documents) contains sufficient text for reliable analysis, with 43% classified as "Good" quality with substantial content exceeding 500 words. The 806 "Thin" documents — containing 100–499 words — provide enough context for basic scoring but may lack the detail needed to assess more nuanced implementation features. The 462 "Stub" entries, containing fewer than 100 words, typically represent brief announcements or press releases that offer minimal substantive content. While we include these in corpus statistics, they contribute little to the analytical results. The total extracted corpus contains 11.4 million words, with a median document length of 1,247 words (IQR: 318–4,892), indicating that a typical AI governance policy provides several pages of substantive content suitable for detailed assessment.

4.1.4 Enriched Corpus

The retrieval and extraction pipeline produced a unified corpus file (`corpus_enriched.json`) that merges OECD metadata with our extracted content and quality assessments. For each of the 2,216 entries, this file preserves the original OECD metadata — including title, jurisdiction, year, URL, policy type, and target sectors — while adding the extracted full text (or OECD snippet where full text was unavailable), text quality classification, word count, and extraction method employed. This enriched structure enables analyses that link policy content to contextual metadata, supporting questions about how governance quality varies by jurisdiction, year, or policy type.

4.1.5 Country Metadata

To enable cross-national comparison, each jurisdiction was mapped to standardized contextual metadata using World Bank classifications. Income groups follow the World Bank's four-tier system: High Income (HI), Upper Middle Income (UMI), Lower Middle Income (LMI), and Low Income (LI). For analyses focused on the North–South divide, we constructed a binary classification contrasting High Income countries against Developing countries (aggregating UMI, LMI, and LI). Regional classifications employ the World Bank's geographic taxonomy: East Asia & Pacific (EAP), Europe & Central Asia (ECA), Latin America & Caribbean (LAC), Middle East & North Africa

(MENA), North America (NAM), South Asia (SA), and Sub-Saharan Africa (SSA). We also incorporated GDP per capita (current US dollars, 2023) as a continuous measure of economic development, enabling analyses that examine governance quality relative to national wealth.

International organisations — including the OECD itself, the European Union, the United Nations, and multilateral development banks — were flagged separately and excluded from country-level analyses where appropriate, as these entities operate under different institutional logics than national governments.

4.1.6 Sample Composition

The final analytical sample reflects the OECD.AI Observatory's coverage, which skews toward high-income countries:

Table 4.3: Sample by income group

Income Group	N Policies	%	N Countries
High Income	1,700	76.7%	~40
Developing	397	17.9%	~30
International	119	5.4%	—
Total	**2,216**	**100%**	**70+**

Table 4.3 reveals a substantial compositional imbalance: high-income countries account for 77% of policies in the corpus, while developing countries contribute only 18%. This disparity reflects the genuine distribution of AI governance activity globally — high-income countries have produced more policies, published more documentation, and maintained more accessible policy archives. However, this imbalance creates analytical challenges, as conventional statistical comparisons assume relatively balanced groups. We address potential selection effects and the implications of unbalanced samples through comprehensive robustness checks in Section 10.1, including analyses restricted to well-documented policies and country-level aggregations that equalize representation.

4.1.7 Analytical Pipeline Overview

The journey from raw OECD.AI metadata to empirical findings involves multiple transformation stages, each addressing distinct methodological challenges. Figure 4.1 visualizes this progression, showing how 2,216 initial entries flow through retrieval, extraction, scoring, and analysis to produce the 120 outputs (figures, tables, statistical tests) that appear in subsequent chapters. This pipeline architecture separates data collection concerns from analytical decisions, enabling transparent documentation of how each methodological choice affects downstream results.

Figure 4.1 shows how each stage transforms the data: from initial policy entries through document retrieval and text extraction (the data collection phase documented in preceding sections), to LLM-based scoring (detailed in Section 5.1), culminating in the 20 analytical chapters that follow. The 6,641 LLM API calls represent three model assessments for each of the 2,216 policies across 10 dimensions, with the ensemble approach ensuring reliability through inter-model agreement.

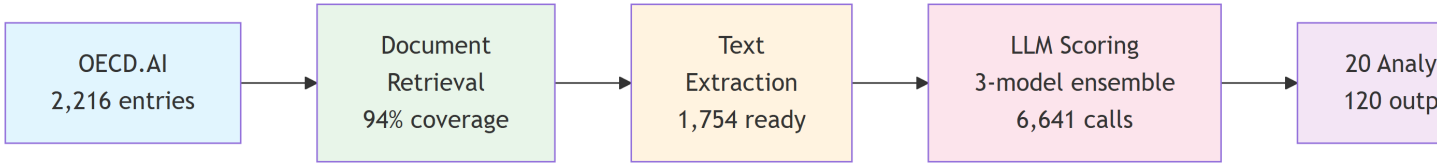

Figure 4.1: Analytical pipeline from corpus to results

4.1.8 Analytical Methods

The statistical analyses in subsequent chapters employ multiple complementary methods to examine governance capacity from different angles. This methodological pluralism enables robust inference: findings that emerge consistently across diverse analytical approaches inspire greater confidence than those dependent on a single modeling choice. Here we overview the core analytical techniques; specific model specifications appear in their respective chapters.

4.1.8.1 Text-to-Data Conversion: LLM Ensemble Scoring

The foundational methodological step — and the innovation that enables analysis at this scale — is the conversion of unstructured policy documents into structured quantitative scores. Unlike traditional text analysis approaches that extract word frequencies, topics, or sentiment, our method employs frontier large language models as expert policy analysts. Each LLM reads the full policy document (up to the model's context window, typically 8,000+ words), applies the detailed scoring rubric for all 10 dimensions simultaneously, and returns structured JSON-formatted scores with textual evidence justifying each assessment. This approach preserves the interpretive sophistication of human expert coding — capturing whether a policy merely mentions implementation features or provides concrete operational details — while achieving the scale necessary to analyze 2,216 documents.

The three-model ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) functions as a panel of expert raters, with the median score serving as the final assessment. This ensemble design addresses the known variability of individual LLM outputs while leveraging their complementary strengths: Claude's nuanced policy interpretation, GPT-4o's balanced analytical approach, and Gemini's efficient processing. The resulting ICC(2,1) = 0.827 demonstrates excellent inter-rater reliability, comparable to or exceeding typical human coder agreement on complex policy dimensions. Detailed validation of this approach, including comparison with human expert ratings, appears in Section 5.1. All subsequent statistical analyses operate on these LLM-derived scores rather than on raw text, treating the scoring outputs as the primary data.

4.1.8.2 Descriptive Analysis

Each analytical chapter begins with descriptive statistics and visual exploration. We present dimension-specific distributions using histogaps (histograms with frequency annotations), ridge plots showing density distributions across groups, and radar charts illustrating multidimensional profiles.

These visualizations reveal patterns that summary statistics alone might obscure — such as bimodality in score distributions or dimension-specific gaps that disappear in composite scores. Box plots with violin overlays show both central tendency and full distributional shape, while heatmaps reveal clustering patterns in policy portfolios across countries and dimensions.

4.1.8.3 Regression Models

Chapters examining determinants of governance capacity employ four complementary regression approaches. Standard OLS regression establishes baseline relationships between predictors (GDP per capita, policy year, document type, text quality) and capacity scores. Multilevel models with random intercepts for countries account for the nested structure of policies within jurisdictions, correcting for dependency that would otherwise inflate standard errors. Quantile regression examines whether predictors affect low-scoring and high-scoring policies differently, revealing heterogeneous effects across the distribution. Tobit models address the substantial floor effect (27.6% of policies score exactly zero) through left-censoring at zero, correcting the attenuation bias that OLS exhibits when floor effects are present.

4.1.8.4 Inequality Analysis

The inequality chapters employ decomposition techniques to partition total variance into meaningful components. Gini coefficients and Lorenz curves quantify overall inequality in governance scores and visualize concentration. Theil's T index enables exact additive decomposition of total inequality into between-group (high-income vs. developing) and within-group components, revealing how much of the apparent North–South divide reflects genuine group differences versus within-group heterogeneity. Policy portfolio analysis examines breadth (whether countries address all dimensions) versus depth (score levels within covered dimensions), distinguishing coverage gaps from implementation quality.

4.1.8.5 Temporal Analysis

Chapters examining governance dynamics over time use panel data methods to separate within-country trends from between-country differences. First-difference models examine year-to-year changes, removing country fixed effects to focus on temporal evolution. We employ Cohen's d effect sizes to assess the substantive significance of changes over time, complementing statistical significance tests that can be misleading with large samples. Convergence analysis tests whether the gap between income groups is narrowing, widening, or remaining stable, using interaction terms between income group and time trends.

4.1.8.6 Multivariate Methods

Principal component analysis (PCA) examines the latent structure underlying the 10 governance dimensions, testing whether capacity and ethics represent empirically distinct constructs. We report eigenvalues, scree plots, and component loadings to assess dimensionality, applying the Kaiser

criterion (eigenvalues > 1) to determine the number of meaningful components. Cronbach's alpha assesses internal consistency of the capacity and ethics subscales, quantifying whether dimensions within each construct reliably measure a coherent latent variable. K-means clustering identifies natural groupings of policies based on their multidimensional profiles, with optimal k determined through silhouette coefficients and stability analysis across bootstrap samples.

4.1.8.7 Hypothesis Testing

Throughout the analyses, we employ both parametric and non-parametric hypothesis tests depending on distributional assumptions. Welch's t-tests compare mean scores between income groups, using the Welch correction to avoid assuming equal variances. Mann-Whitney U tests provide non-parametric alternatives when distributions violate normality assumptions. Chi-square tests assess whether categorical outcomes (such as quadrant membership in the capacity–ethics space) differ by income group. For all tests, we report exact p-values, effect sizes (Cohen's d for mean differences, Cramér's V for categorical associations), and confidence intervals where appropriate, following contemporary standards for transparent statistical reporting.

4.1.9 Reproducibility

All code is available at https://github.com/lsempe77/ai-governance-capacity. The pipeline uses deterministic document IDs (`MD5(url)[:12]`) to ensure reproducibility of the corpus-to-analysis link. API calls to LLM providers used fixed model identifiers and structured JSON output schemas.

4.1.10 Use of Large Language Models

This research employs large language models in two distinct capacities, both of which we disclose here in the interest of methodological transparency.

For data analysis: Large language models (Claude Sonnet 4, GPT-4o, and Gemini Flash 2.0) serve as the core analytical instrument, functioning as automated policy coders that convert unstructured policy documents into structured quantitative scores. This use constitutes the research methodology itself and is documented extensively throughout Section 4.1 and Section 5.1, including validation against human expert ratings. All LLM-generated scores are preserved in the public repository, enabling verification and replication of our analytical pipeline.

For writing assistance: Large language models (primarily GitHub Copilot and Claude) provided assistance with text editing during manuscript preparation. All LLM-generated text was reviewed, revised, and approved by the author, who takes full responsibility for the accuracy and integrity of the final content. LLMs did not generate substantive intellectual contributions, interpret findings, or make analytical decisions — these remained under direct human control throughout the research process.

This dual disclosure reflects our commitment to transparency in an era where LLM use in research is becoming ubiquitous. We distinguish between LLMs as research instruments (where their use is

the methodology being validated) and LLMs as writing assistants (where they augment but do not replace human scholarly judgment).

5 LLM Ensemble Scoring & Validation

5.1 Measuring Governance Quality at Scale

> **Chapter summary.** This chapter presents our LLM-based scoring methodology — a three-model ensemble that independently codes each policy on 10 dimensions. We report inter-rater reliability (ICC = 0.827, Excellent) and discuss model-specific scoring patterns.

5.1.1 Scoring Framework

The transition from collected documents to analyzable data required developing a comprehensive assessment framework that could systematically evaluate implementation readiness across diverse policy types, jurisdictions, and governance traditions. This framework needed to capture both the structural features that enable implementation (capacity dimensions) and the substantive ethical commitments that shape governance outcomes (ethics dimensions). Drawing on decades of implementation science and the emerging AI governance literature, we constructed a 10-dimension assessment framework organized into two complementary domains.

Each of the 2,216 policies was scored on **10 dimensions** using a 0–4 scale, where 0 indicates complete absence of the feature, 1-2 represent minimal to moderate presence, 3 indicates substantial implementation readiness, and 4 reflects comprehensive operationalization with concrete mechanisms. This five-point scale provides sufficient granularity to distinguish meaningful quality differences while maintaining inter-rater reliability — finer scales would introduce excessive noise, while coarser scales would obscure important variation.

5.1.1.1 Capacity Dimensions

Grounded in implementation science (Mazmanian and Sabatier 1983; Lipsky 1980; Grindle 1996; Fukuyama 2013):

Table 5.1: Capacity scoring dimensions

Code	Dimension	What It Measures
C1	Clarity & Specificity	Clear objectives, measurable targets, defined scope
C2	Resources & Budget	Dedicated funding, staffing, infrastructure

Code	Dimension	What It Measures
C3	Authority & Enforcement	Legal mandate, penalties, compliance mechanisms
C4	Accountability & M&E	Reporting, evaluation, oversight bodies
C5	Coherence & Coordination	Cross-agency alignment, international coordination

These five capacity dimensions operationalize the implementation conditions identified by Mazmanian and Sabatier (1983) and extended by subsequent scholars. Clarity corresponds to Mazmanian and Sabatier's emphasis on clear objectives and causal theories; Resources captures Grindle's technical and fiscal capacity requirements; Authority reflects the legal structuring of implementation processes; Accountability operationalizes Lipsky's concern with constraining street-level discretion; and Coherence addresses the coordination challenges documented by Hjern and Hull (1982). Together, they provide a comprehensive assessment of whether policies possess the institutional infrastructure necessary for execution.

5.1.1.2 Ethics Dimensions

Grounded in AI ethics literature (Jobin, Ienca, and Vayena 2019; Floridi et al. 2018; OECD 2019; UNESCO 2021; European Parliament and Council 2024):

Table 5.2: Ethics scoring dimensions

Code	Dimension	What It Measures
E1	Ethical Framework Depth	Grounding in principles, coherent ethical vision
E2	Rights Protection	Privacy, non-discrimination, human oversight, transparency
E3	Governance Mechanisms	Ethics boards, impact assessments, auditing
E4	Operationalisation	Concrete requirements, standards, certification
E5	Inclusion & Participation	Stakeholder processes, marginalised group representation

The ethics dimensions synthesize principles identified across the AI governance literature, particularly the convergence documented by Jobin, Ienca, and Vayena (2019) around transparency, fairness, accountability, and privacy. Framework Depth assesses whether policies ground specific requirements in coherent ethical visions rather than listing buzzwords. Rights Protection operationalizes the human-centric principles emphasized by Floridi et al. (2018) and enshrined in frameworks like UNESCO's AI Recommendation. Governance Mechanisms capture the institutional architecture

for ethics oversight, while Operationalisation distinguishes aspirational statements from concrete requirements with measurable standards. Inclusion reflects the participatory governance emphasis in OECD (2019), recognizing that AI governance legitimacy depends on meaningful stakeholder engagement.

Each dimension uses explicit scoring rubrics (see Section A.1) with anchored examples at each scale point, ensuring that assessments rest on observable textual evidence rather than subjective impressions. Composite scores are computed as unweighted means: *Capacity* = mean(C1–C5), *Ethics* = mean(E1–E5), *Overall* = mean(all 10). This equal weighting reflects our agnostic stance on which dimensions matter most — different governance contexts may prioritize different features, and our framework captures this multidimensionality rather than imposing a single definition of quality.

5.1.2 Three-Model Ensemble

Applying this 10-dimension framework to 2,216 documents requires a scoring approach that balances three competing demands: analytical sophistication (capturing nuanced implementation features), scale (processing millions of words of policy text), and reliability (producing consistent assessments across documents). Traditional human expert coding offers sophistication but becomes prohibitively expensive and time-consuming at this corpus size. Automated keyword-based approaches scale efficiently but lack the interpretive capacity to distinguish substantive implementation details from aspirational rhetoric. Our solution employs frontier large language models as automated policy analysts, leveraging their ability to read and interpret complex documents while maintaining consistency through ensemble design.

To mitigate single-model bias and architectural idiosyncrasies, each policy was independently scored by three frontier LLMs via the OpenRouter API, selected to represent diverse training approaches and institutional origins:

Table 5.3: LLM ensemble composition

Model	Identifier	Role	Entries Scored
Model A	Claude Sonnet 4	Strictest scorer	2,210 (99.7%)
Model B	GPT-4o	Moderate scorer	2,216 (100%)
Model C	Gemini Flash 2.0	Moderate scorer	2,215 (100%)

This ensemble design leverages complementary strengths: Claude Sonnet 4's nuanced policy interpretation and attention to implementation details, GPT-4o's balanced analytical approach and broad domain knowledge, and Gemini Flash 2.0's efficient processing and consistent scoring patterns. By combining models from three different organizations (Anthropic, OpenAI, Google) trained on potentially different corpora using different architectures, we reduce the risk that shared training biases or architectural quirks systematically skew results.

Each model received identical structured prompts containing the full policy text (up to context window limits, typically 8,000+ words) and the complete scoring rubric with anchored examples. The prompts instructed models to read the entire document, assess each dimension independently,

assign a 0-4 score based on observable textual evidence, and provide brief supporting excerpts justifying each score. Models returned structured JSON-formatted outputs with dimension-level scores and evidence, enabling automated aggregation while preserving auditability through the evidence field. The final ensemble score for each dimension is the **median** of the three model scores, following the logic of robust central tendency estimation. The median approach proves superior to the mean in this context because it remains unaffected by single-model outliers and handles the systematic calibration differences we observe across models (detailed below) without requiring explicit recalibration.

The total scoring effort required **6,641 API calls** (2,216 policies × 3 models, minus a handful of failures where models returned malformed JSON or exceeded context windows). The high completion rate — 99.7% of entries successfully scored by all three models — demonstrates the robustness of the pipeline to diverse document formats and lengths.

5.1.3 Inter-Rater Reliability

The validity of this entire analytical enterprise rests on a fundamental question: do the three models agree on policy quality, or do they produce idiosyncratic assessments that reflect model-specific biases rather than genuine document features? If inter-model agreement is low, the ensemble scores become arbitrary — different model combinations would yield different conclusions. If agreement is high, this provides evidence that the scores capture systematic variation in policy quality rather than measurement noise.

We assess agreement across the three LLM "raters" using multiple complementary metrics, following the framework established by Shrout and Fleiss (1979) for inter-rater reliability in observational studies. The intraclass correlation coefficient ICC(2,1) serves as our primary reliability measure, as it appropriately handles the nested structure of our data (three models rating each policy) and quantifies the proportion of total variance attributable to true between-policy differences rather than rater disagreement. We supplement this with pairwise correlations, Fleiss' kappa for categorical agreement, and descriptive measures of score spread to provide a comprehensive reliability portrait.

5.1.3.1 Overall Reliability

Table 5.4: Inter-rater reliability summary

Metric	Value	Interpretation
ICC(2,1) overall	**0.827**	Excellent
ICC(2,1) capacity	0.824	Excellent
ICC(2,1) ethics	0.791	Excellent
Mean pairwise Pearson	0.86	Strong
Mean pairwise Spearman	0.88	Strong
Mean Fleiss'	0.51	Moderate
Mean overall spread	0.40/4	Low disagreement
Scores within 1 point	95.4%	High consistency

Metric	Value	Interpretation

Table 5.4 presents a remarkably consistent picture across multiple metrics. The ICC(2,1) of 0.827 indicates "Excellent" reliability under Cicchetti's (1994) guidelines (>0.75 = Excellent), meaning that approximately 83% of the variance in observed scores reflects true differences between policies rather than rater disagreement. This level of agreement is comparable to or exceeds reliability typically reported in human-coded policy analysis studies, where ICC values of 0.70-0.80 are considered strong evidence of coding quality. The high pairwise correlations (mean r = 0.86, = 0.88) confirm this consistency through a different lens, while the low mean spread (0.40 points on a 4-point scale) and high within-1-point agreement (95.4%) demonstrate that models rarely produce wildly divergent assessments. Even Fleiss' kappa — a more conservative metric that treats the 0-4 scale categorically rather than continuously — achieves moderate agreement (0.51), which for a five-category scale represents substantial consensus.

Crucially, both capacity and ethics subscales achieve excellent reliability independently (ICC = 0.824 and 0.791 respectively), indicating that the strong overall agreement is not driven by a single dominant construct but reflects genuine consensus across both theoretical domains.

5.1.3.2 Dimension-Level ICCs

Table 5.5: Dimension-level ICC values

Dimension	ICC(2,1)	Quality
C1 Clarity	0.720	Good
C2 Resources	0.735	Good
C3 Authority	0.751	Excellent
C4 Accountability	0.753	Excellent
C5 Coherence	0.804	Excellent
E1 Framework	0.751	Excellent
E2 Rights	0.785	Excellent
E3 Governance	0.691	Good
E4 Operationalisation	0.605	Good
E5 Inclusion	0.746	Good

Table 5.5 reveals systematic patterns in dimension-level reliability that illuminate the scoring process. All dimensions achieve at least "Good" reliability (>0.60), with six reaching "Excellent" (>0.75). The highest agreement appears on structural features like Coherence (ICC = 0.804), Authority (0.751), and Rights Protection (0.785) — dimensions where textual evidence is relatively concrete and unambiguous. Lower (though still acceptable) reliability on Operationalisation (0.605) and Governance Mechanisms (0.691) likely reflects the greater interpretive challenge these dimensions pose: distinguishing truly operational requirements from aspirational language requires subtle judgment that even sophisticated models may approach differently. The lowest ICC (E4 Operationalisation, 0.605) still comfortably exceeds conventional acceptability thresholds (>0.40

for exploratory research, >0.60 for established scales), providing confidence that all 10 dimensions contribute meaningful signal rather than noise to the composite scores.

5.1.3.3 Model-Specific Scoring Patterns

The three models exhibit systematic scoring tendencies:

Table 5.6: Model-level mean scores

Model	Capacity Mean	Ethics Mean	Overall Mean
A (Claude)	0.68	0.46	0.57
B (GPT-4o)	0.92	0.71	0.81
C (Gemini)	0.93	0.68	0.81

Table 5.6 exposes a striking and systematic pattern: Model A (Claude Sonnet 4) scores approximately 0.24 points lower on average than Models B and C across both capacity and ethics dimensions. This is not random noise or jurisdiction-specific bias — the pattern holds consistently across all policy types, income groups, and regions, indicating a fundamental calibration difference in how the model interprets the 0-4 scale. Model A appears to require stronger textual evidence to assign higher scores, treating the rubric descriptions more stringently than its counterparts. The gap is particularly pronounced on ethics dimensions (0.46 vs. 0.68-0.71), suggesting that Model A applies more demanding standards for what constitutes operationalized ethical governance versus aspirational principles.

Importantly, this systematic shift does not invalidate Model A's contributions to the ensemble. The high correlation between Model A's scores and those of Models B and C ($r > 0.85$) demonstrates that all three models agree on the *rank ordering* of policies even while disagreeing on absolute levels. The median-based aggregation proves robust to this calibration difference: it preserves the relative rankings while positioning the final scores between the strict and lenient interpretations. An alternative approach using mean scores would require explicit recalibration or standardization; the median avoids this complexity while naturally accounting for systematic shifts.

5.1.3.4 Agreement by Text Quality

Table 5.7: Agreement by text quality

Text Quality	N	Mean Spread	Within 1 pt
Good (500 words)	942	0.57	90.3%
Thin (100–499)	805	0.34	98.9%
Stub (<100)	462	0.13	99.8%

Table 5.7 reveals the expected relationship between document informativeness and scoring consensus. Models achieve near-perfect agreement on stub documents (mean spread 0.13, within-1-point agreement 99.8%), largely because these minimal texts provide insufficient evidence for any dimension to score above zero. The models converge trivially on low scores when documents offer little substance to assess. Agreement remains very high on thin documents (spread 0.34, agreement 98.9%), as these 100-499 word texts typically mention governance features without providing implementation details, again limiting the interpretive range.

The elevated disagreement on good-quality texts (spread 0.57, agreement 90.3%) should not be interpreted as a reliability failure but rather as evidence that models are engaging substantively with document content. Longer, more detailed policies present genuinely ambiguous cases where reasonable analysts might differ: Does a policy with detailed budget projections but unclear enforcement mechanisms score 2 or 3 on Resources? Does sophisticated ethical framework discussion without concrete operationalization merit a 2 or 3 on Framework Depth? These interpretive challenges produce the higher spread we observe. The fact that even for good texts, 90.3% of scores fall within 1 point indicates that disagreement occurs at boundary cases rather than reflecting fundamental divergence in assessment.

5.1.4 Composite Scores

The resulting ensemble produces composite scores with the following distributions:

Table 5.8: Composite score distributions

Component	Mean	SD	Median	IQR
Capacity (C1–C5)	0.83	0.77	0.60	0.00–1.40
Ethics (E1–E5)	0.61	0.62	0.40	0.00–1.00
Overall (all 10)	0.73	0.66	0.50	0.10–1.15

Table 5.8 summarizes the final ensemble scores that serve as the primary data for all subsequent analyses. Three distributional features prove particularly consequential for analytical choices in later chapters.

First, the **strong floor effect** — with 27.6% of policies scoring exactly zero on capacity and 36.3% on ethics — indicates that more than a quarter of documents in the OECD.AI Observatory contain insufficient implementation detail to score above the minimum threshold on our framework. These zeros are not missing data but substantive findings: many AI governance documents consist of brief announcements, aspirational statements, or high-level principles without operational content. This censoring at zero violates the assumptions of standard OLS regression, motivating the Tobit models we employ in **?@sec-cap-determinants** to correct for attenuation bias.

Second, the **right skew** in all three distributions — with medians substantially below means and interquartile ranges concentrated in the lower half of the scale — reveals that most policies cluster at the low end of implementation readiness, while a smaller set of comprehensive policies achieve substantially higher scores. This heterogeneity suggests that focusing solely on mean comparisons

would obscure important distributional differences, motivating the quantile regression approach that examines effects at different points of the score distribution.

Third, the systematic **capacity-ethics gap** — with policies averaging 0.83 on implementation architecture but only 0.61 on ethics operationalization — points to a prioritization pattern: governments more frequently specify institutional structures, budgets, and authorities than operationalize ethical principles through concrete requirements. This gap receives detailed examination in **?@sec-pca-nexus**, where we explore the capacity-ethics nexus and identify distinct governance typologies.

5.1.5 Validation Discussion

The use of large language models as automated policy coders represents a methodological innovation with both promise and peril. Our approach builds on a growing body of evidence demonstrating that frontier language models can perform complex text annotation tasks at or above human-coder quality (Gilardi, Alizadeh, and Kubli 2023; TÃ¶rnberg 2024). Recent validation studies show that LLMs achieve reliability comparable to trained human coders on tasks ranging from sentiment classification to ideological scaling, while processing text orders of magnitude faster and at far lower cost. However, these findings come with important caveats (Pangakis, Wolken, and Fasching 2023): LLM performance varies substantially across task types, prompt formulations, and model versions, and models can exhibit systematic biases learned from training data that may not align with human expert judgment on normatively contentious dimensions.

Three features of our methodological design directly address these validity concerns. The **multi-model ensemble** reduces the risk that findings reflect idiosyncrasies of any single model's training data or architectural choices by combining three independently-developed models from different organizations. If all three models converge on similar assessments despite their different origins, this provides stronger evidence of validity than relying on a single model's output. The **structured output with evidence** requirement — where models must provide supporting textual excerpts justifying each score — enables post-hoc auditing and increases the probability that models ground assessments in observable document features rather than generating plausible-sounding scores without textual basis. The **median aggregation** strategy proves robust both to single-model outliers and to the systematic calibration difference we observe across models, avoiding the need for explicit recalibration while preserving relative rankings.

Important limitations remain that readers should bear in mind when interpreting results. The three models, despite their different origins, may share biases inherited from overlapping training corpora — particularly given that all were likely exposed to prominent AI governance documents like the OECD AI Principles and EU AI Act during training. The scoring rubric itself, while grounded in implementation science theory and AI governance scholarship, necessarily involves subjective judgments about what constitutes "adequate" clarity or "substantial" resource allocation — dimensions on which even expert human coders would reasonably disagree. Our ensemble treats all three models as equally authoritative through median aggregation, but this may not reflect their actual relative validity — it is conceivable that one model's systematic stringency or leniency better aligns with ground truth than the ensemble median, though we lack a gold standard against which to evaluate this.

These methodological uncertainties motivate the extensive robustness checks presented in Section 10.1, where we examine whether core findings hold across alternative specifications, subsamples, and aggregation methods. The consistency of results across these checks provides additional confidence that our conclusions reflect genuine patterns in policy quality rather than artifacts of measurement choices.

6 UNESCO Alignment Landscape

6.1 The Alignment Landscape: Coverage and Depth

Chapter summary. This chapter maps how well 1,326 AI policies from 79 jurisdictions address the 25 items of the UNESCO Recommendation on the Ethics of Artificial Intelligence. Mean alignment is moderate ($\mu = 53.9$, $\sigma = 12.2$), with a clear "implementation gap": values (55%) and principles (53%) are better covered than policy action areas (41%). The most-addressed item — peaceful and just societies (83%) — is not the most substantively developed, revealing a "paradox of proclamation" in which breadth of mention does not guarantee depth of engagement.

6.1.1 Overall Alignment Score Distribution

We define a composite **UNESCO alignment score** (0–100) for each policy, weighting coverage breadth (60%) and normalised depth quality (40%) across the 25 UNESCO items — 4 values, 10 principles, and 11 policy action areas. The score captures both *whether* a policy mentions a UNESCO item and *how substantively* it engages with it.

The mean alignment score is **53.9** (SD = 12.2, median = 53.8). No policy achieves full alignment: the maximum score is approximately 85, and the minimum near 20. This moderate central tendency suggests that most policies engage with roughly half of the UNESCO framework at a moderate depth — a glass-half-full finding that nonetheless leaves substantial room for improvement.

6.1.2 Coverage Across the 25 UNESCO Items

The coverage landscape is strikingly uneven. The five most-covered items are:

Table 6.1: Most-addressed UNESCO items

UNESCO Item	Type	Coverage
Peaceful, just & interconnected societies	Value	83.4%
Ethical governance & stewardship	Policy area	82.4%
Economy & labour	Policy area	79.6%
Responsibility & accountability	Principle	72.8%
Fairness & non-discrimination	Principle	72.5%

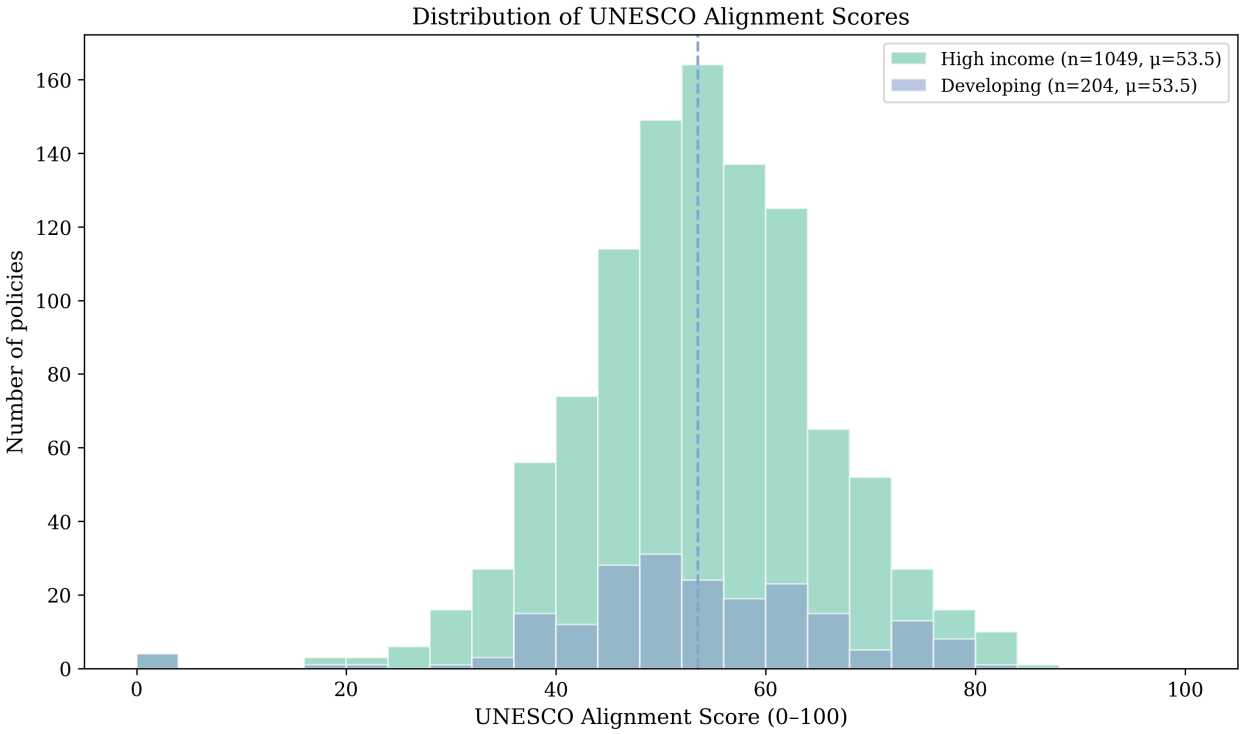

Figure 6.1: Distribution of UNESCO alignment scores across 1,326 AI policies. The distribution is approximately normal, centred on 54 with moderate spread.

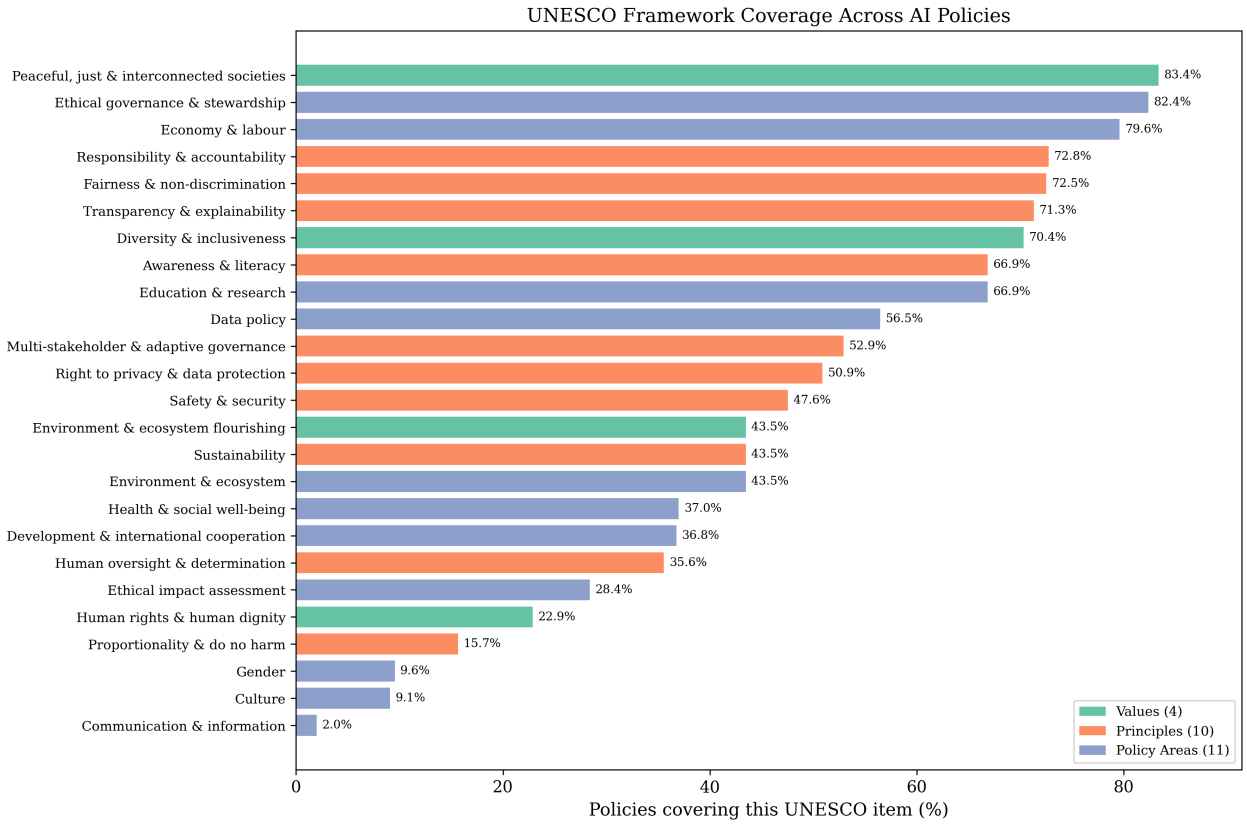

Figure 6.2: Coverage rates across all 25 UNESCO items, ordered by frequency. Five items exceed 70% coverage, while six fall below 30%.

By contrast, the five least-covered items are:

Table 6.2: Least-addressed UNESCO items

UNESCO Item	Type	Coverage
Communication & information	Policy area	2.0%
Culture	Policy area	9.1%
Gender	Policy area	9.6%
Proportionality & do no harm	Principle	15.7%
Human rights & human dignity	Value	22.9%

The **near-absence of communication and information** (2.0%) represents a genuine blind spot: despite the UNESCO Recommendation explicitly calling for policies to address AI's impact on media, information ecosystems, and freedom of expression, virtually no national policy does so. Similarly, **human rights and human dignity** — the foundational value of the Recommendation and arguably its normative anchor — appears in only **22.9%** of policies. This is a striking omission for a framework that explicitly grounds AI ethics in human rights.

6.1.3 The Implementation Gap: Values vs. Principles vs. Policy Areas

The UNESCO Recommendation's three-tier structure reveals a gradient of decreasing policy engagement:

Table 6.3: Coverage by UNESCO framework layer

Layer	Mean Coverage	Items
Values	55.0%	4
Principles	53.0%	10
Policy action areas	41.1%	11

This gradient — **values > principles > policy areas** — constitutes what we term the **implementation gap**. Policies are more likely to declare broad ethical values than to specify the concrete governance mechanisms needed to operationalise them. The 14-percentage-point gap between values and policy areas echoes the broader "principles-to-practice" critique in the AI ethics literature.

6.1.4 Coverage vs. Depth: The Paradox of Proclamation

A central finding is that **coverage does not predict depth** ($r = 0.02$, $p = 0.94$). Items that are mentioned by many policies are not necessarily treated more substantively. We term this the **paradox of proclamation**: the most frequently invoked principles (transparency, accountability, fairness) tend to appear as brief rhetorical gestures — a sentence or even a phrase — rather than as deeply developed policy commitments.

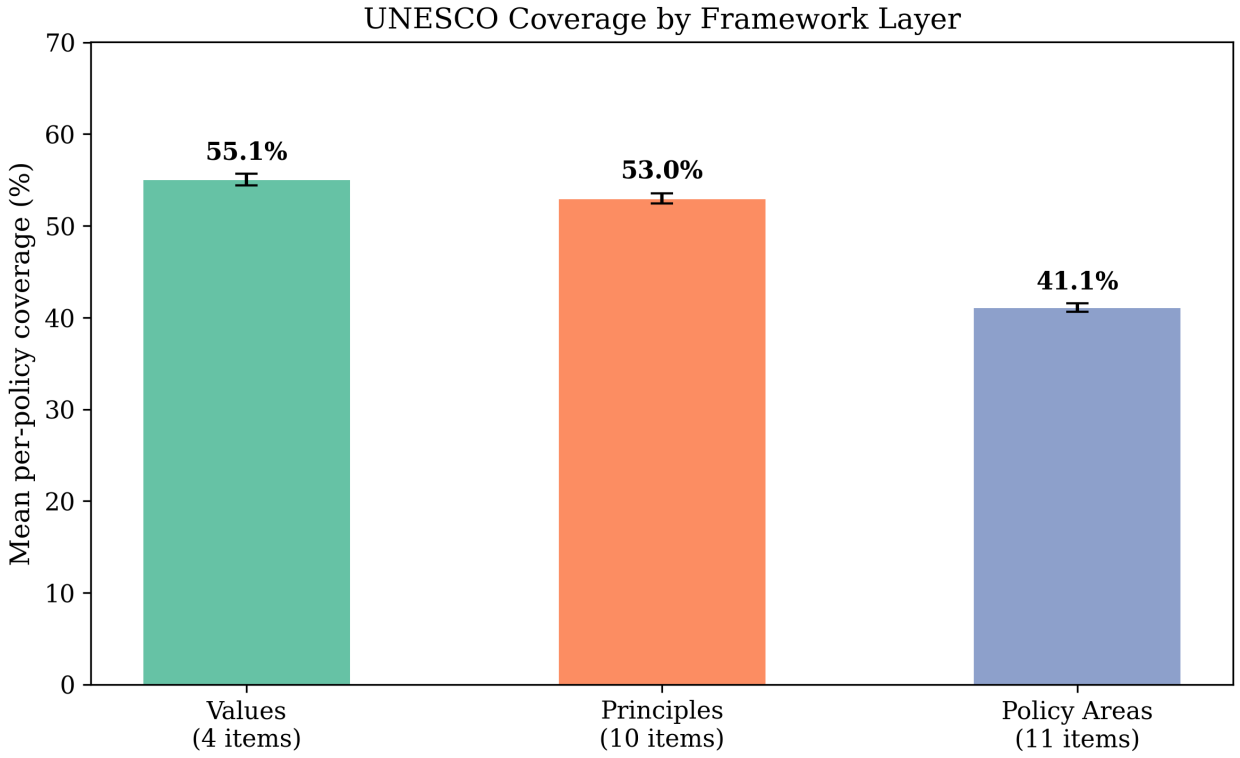

Figure 6.3: Mean coverage rates by UNESCO framework layer. Values and principles are better covered than policy action areas, revealing an "implementation gap."

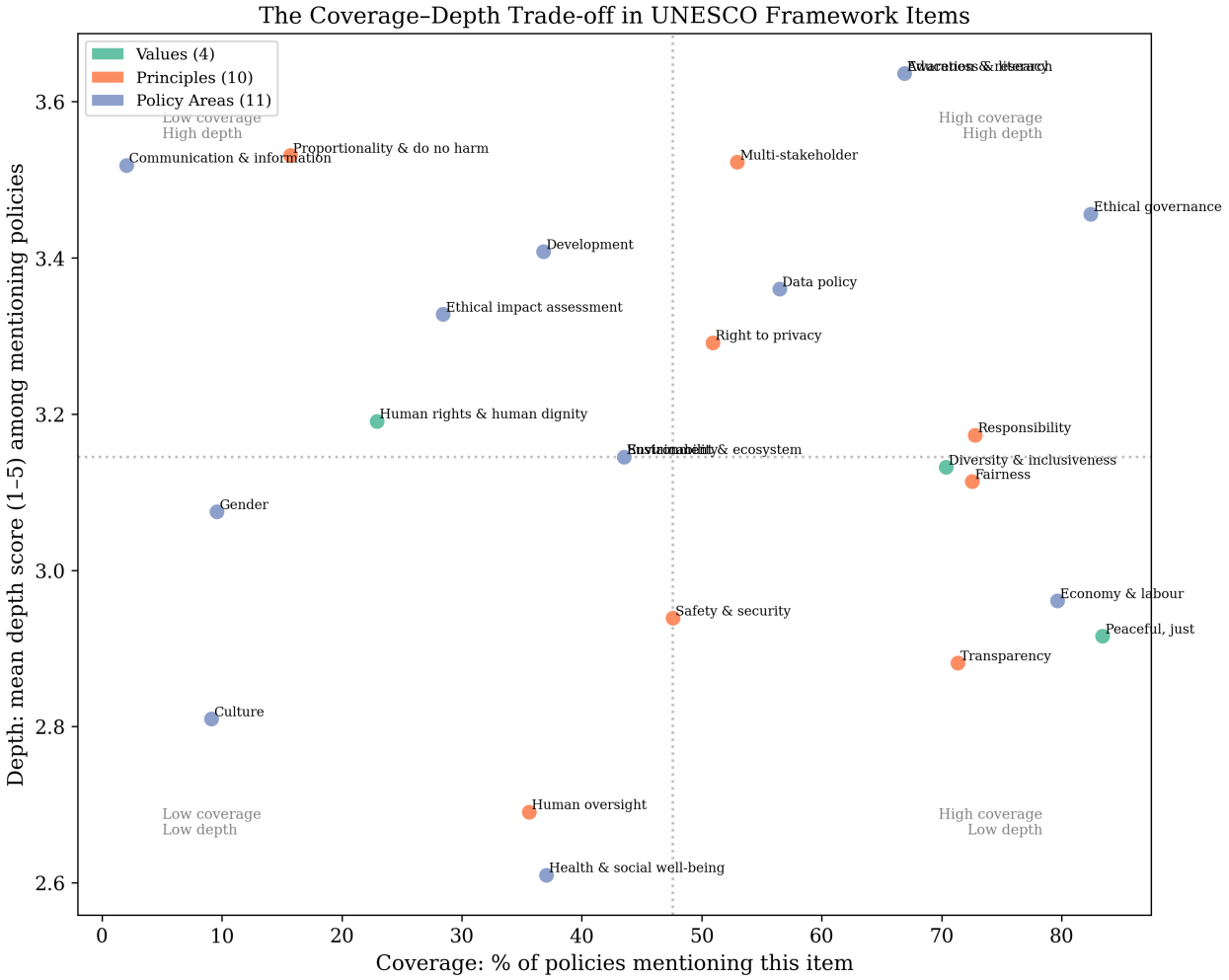

Figure 6.4: Scatter plot of coverage (% of policies mentioning an item) vs. mean depth (1–5 scale) for each UNESCO item. There is no significant correlation ($r = 0.02$, $p = 0.94$), indicating that breadth of mention does not predict substantive engagement.

Conversely, some rarely mentioned items, when they do appear, receive surprisingly substantive treatment. **Awareness and literacy**, for instance, is covered in 66.9% of policies with a mean depth of **3.64** (paragraph-level), reflecting the concrete nature of education programmes. **Proportionality and do no harm**, though covered by only 15.7% of policies, achieves a depth of **3.53** when it does appear — suggesting that the few policies that engage with this principle do so thoughtfully.

6.1.5 The Depth Heatmap

The depth heatmap reveals that most UNESCO items, when mentioned, receive **sentence-level (3)** to **paragraph-level (4)** treatment. Very few items are engaged at section-level (5), suggesting that even the most substantive policies rarely dedicate entire sections to individual UNESCO items.

6.1.6 Depth by Framework Layer

Across all three layers, the depth distributions are roughly similar, with most engagement occurring at depth levels 3–4. However, **policy action areas** show slightly higher depth when present — likely because they translate into concrete programmatic commitments (education, data governance, health) that inherently require more detailed treatment than abstract values.

6.1.7 The Gaps: What UNESCO Asks For and What It Gets

The gap analysis reveals three categories of UNESCO alignment:

1. **Well-integrated items** (>70% coverage): peaceful societies, ethical governance, economy and labour, responsibility, fairness, transparency, diversity. These represent the "consensus core" of global AI ethics — the items that virtually all policy traditions address.
2. **Partially addressed items** (30–70%): data policy, education, privacy, safety, sustainability, environment, health, development cooperation, multi-stakeholder governance, awareness. These are engaged by a majority or near-majority of policies but with significant variation.
3. **Systematically neglected items** (<30%): human rights & dignity (22.9%), proportionality (15.7%), ethical impact assessment (28.4%), gender (9.6%), culture (9.1%), communication & information (2.0%). These represent the **most significant misalignment** between the UNESCO Recommendation and actual global AI policy practice.

The neglect of human rights as a foundational frame, and the near-absence of gender-specific and cultural considerations, suggests that the global AI policy landscape remains heavily shaped by a **technology-centric governance paradigm** rather than the rights-based approach the UNESCO Recommendation advocates.

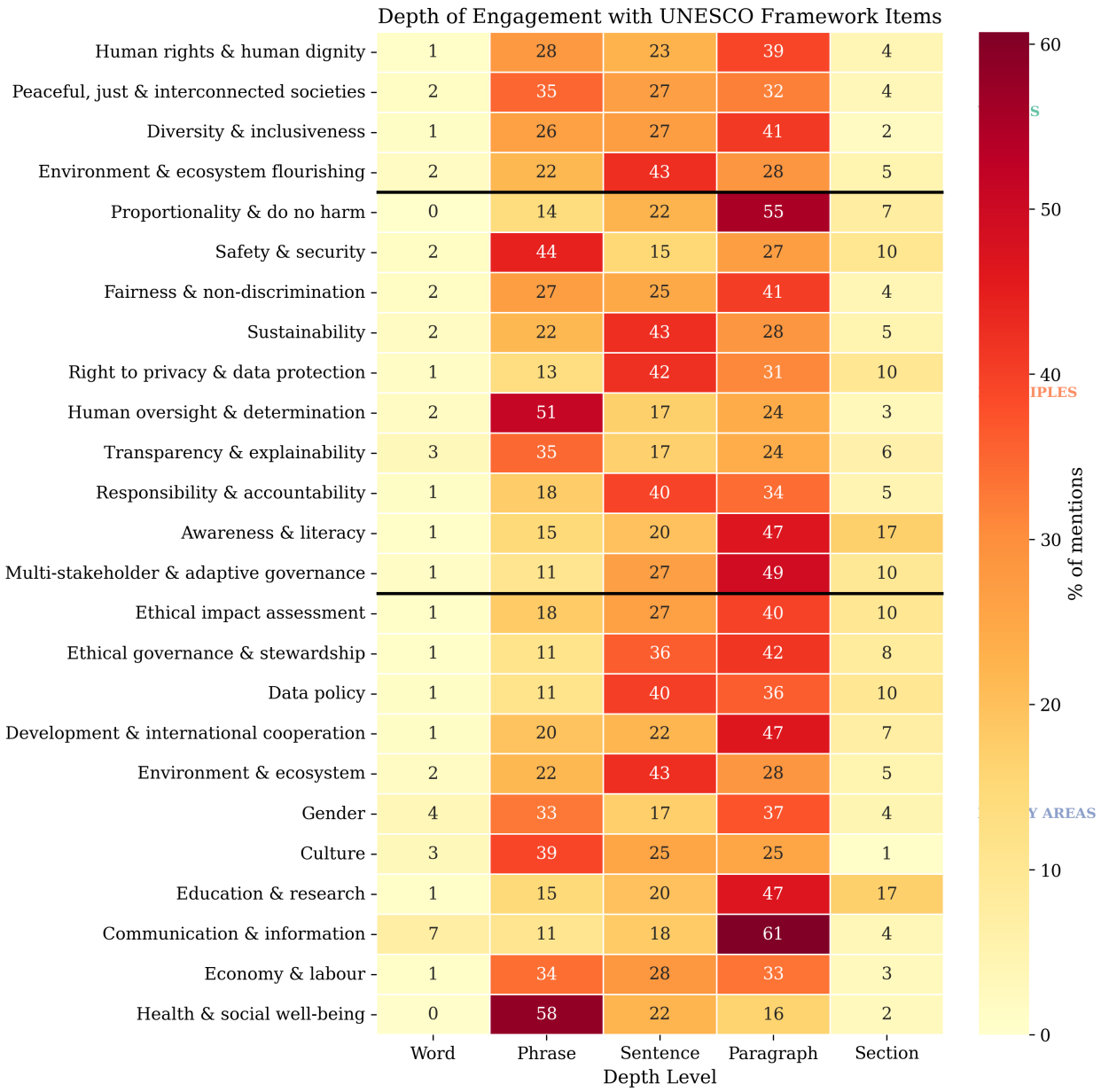

Figure 6.5: Depth heatmap across all 25 UNESCO items, showing the distribution of engagement levels from word (1) to section (5). Most engagement occurs at sentence (3) to paragraph (4) level.

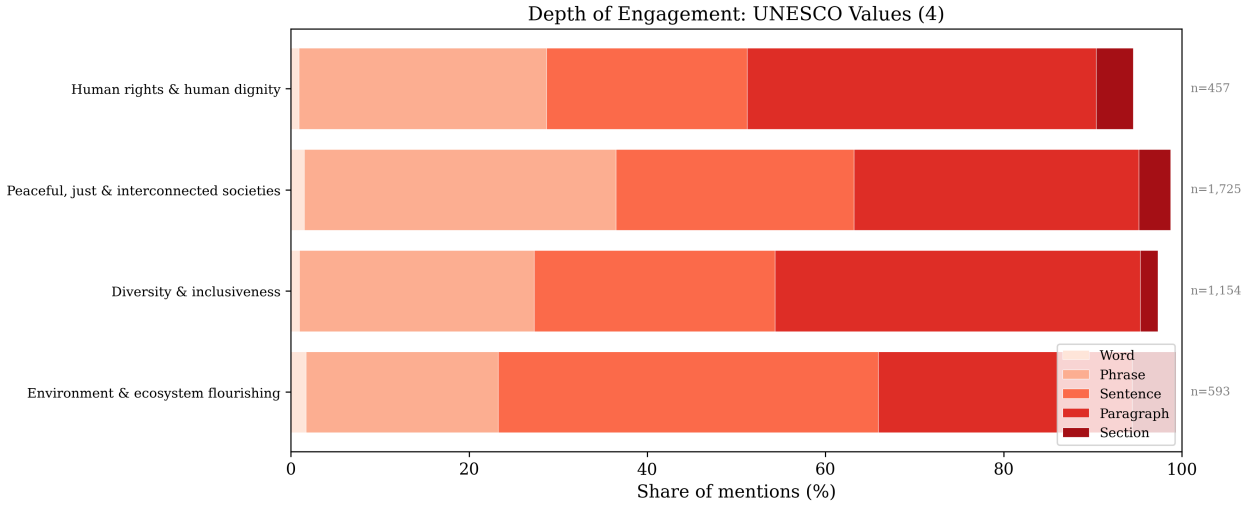

Figure 6.6: Depth distributions for the 4 UNESCO values.

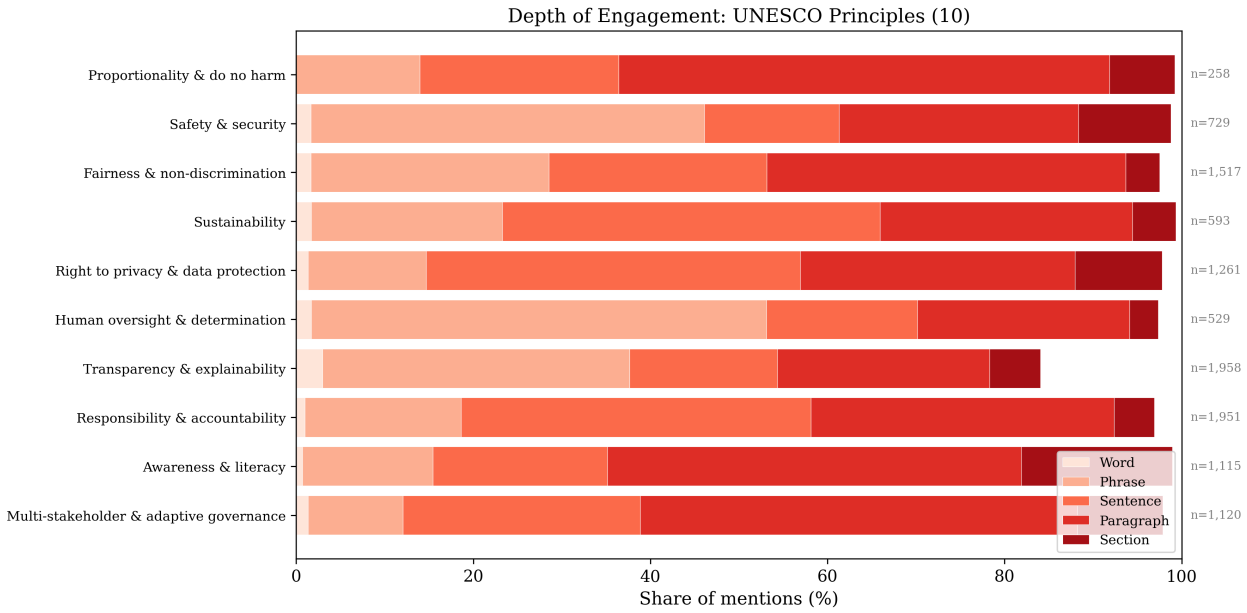

Figure 6.7: Depth distributions for the 10 UNESCO principles.

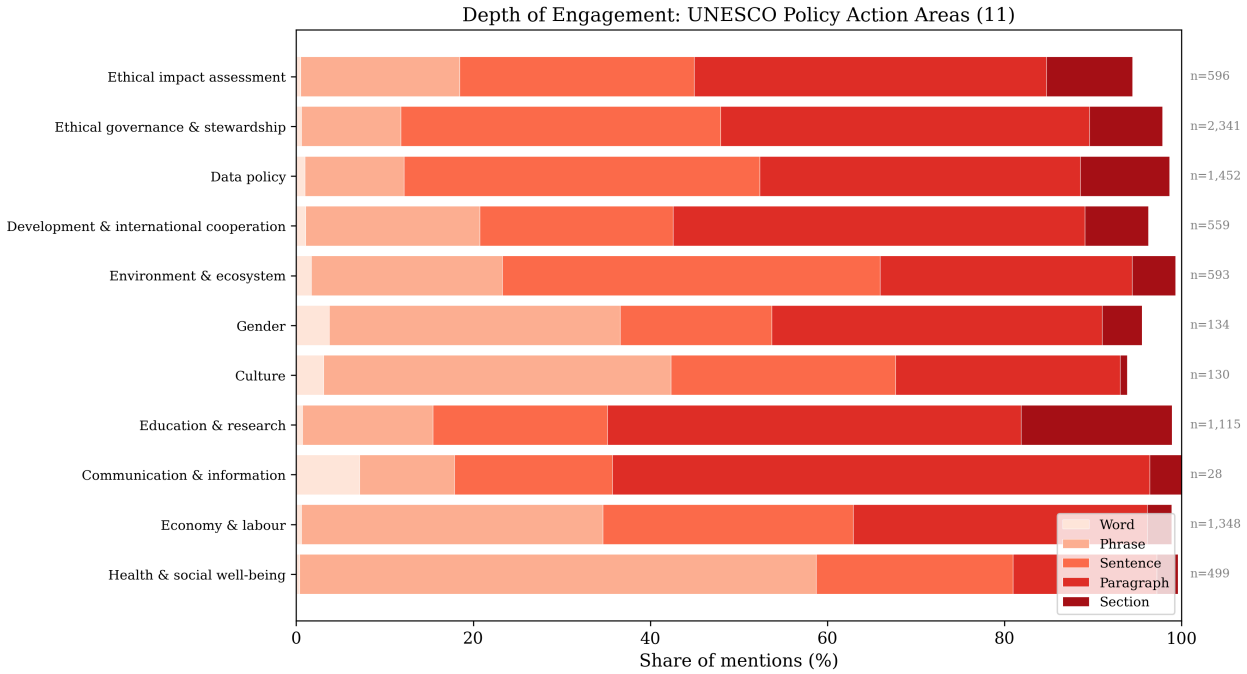

Figure 6.8: Depth distributions for the 11 UNESCO policy action areas.

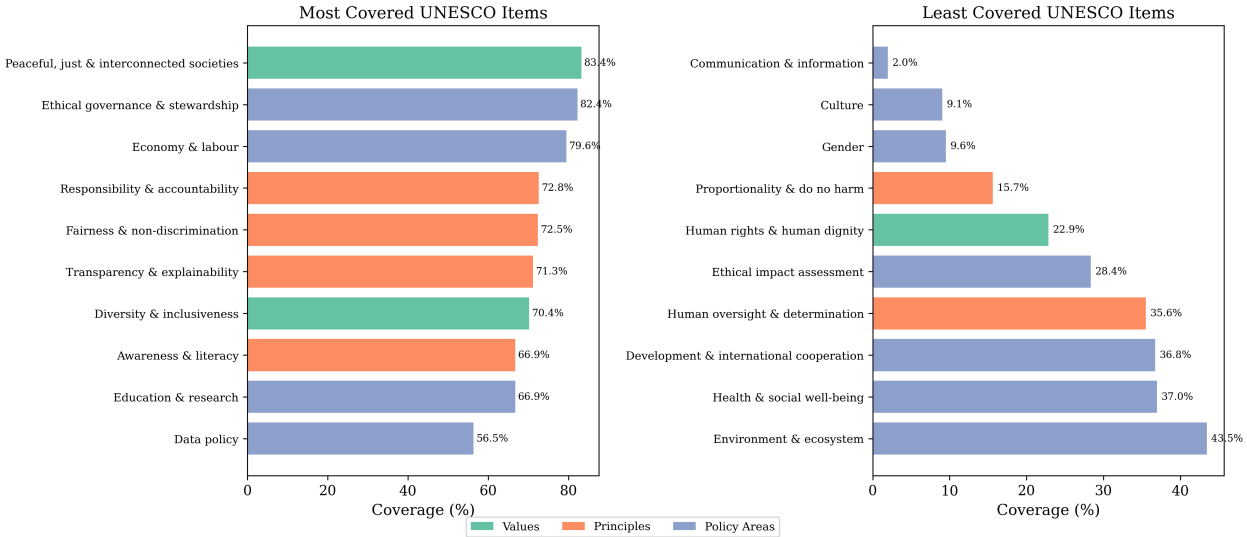

Figure 6.9: Top and bottom UNESCO items by coverage, highlighting the largest gaps between the Recommendation's aspirations and actual policy content.

7 UNESCO Alignment Determinants

7.1 What Drives UNESCO Alignment?

Chapter summary. This chapter examines the structural determinants of UNESCO alignment. The most striking finding is a **null income effect**: high-income and developing countries achieve virtually identical overall alignment scores ($d = 0.001$, $p = 0.99$). Yet this aggregate equivalence conceals significant item-level divergence — developing countries are **more** aligned on health (+19.3pp) and gender (+6.2pp). Regional variation exists but is modest. Soft law instruments score higher than hard law, and a multivariate regression confirms that the strongest predictors of UNESCO alignment are a policy's own **capacity** ($\beta = 3.2$) and **ethics** ($\beta = 11.5$) scores — not national income.

7.1.1 The Income Divide That Isn't

Table 7.1: Income-group comparison for UNESCO alignment

Metric	Value
High income mean (N = 1,049)	53.5
Developing mean (N = 204)	53.5
Welch's t	0.013
p-value	0.99
Cohen's d	**0.001**

The overall income gap in UNESCO alignment is **effectively zero** ($d = 0.001$). This is a striking and counter-intuitive result. The dominant narrative in AI governance discourse — that developing countries lack the institutional capacity to produce sophisticated AI policies — finds no support in our data when alignment is measured against the UNESCO framework.

This null finding is even more remarkable than the small-but-significant gaps observed for capacity ($d = 0.30$) and ethics ($d = 0.20$) scores in Parts I and II. The UNESCO framework, as an internationally negotiated instrument designed with input from all member states, may serve as a normative template that is equally accessible regardless of national income level.

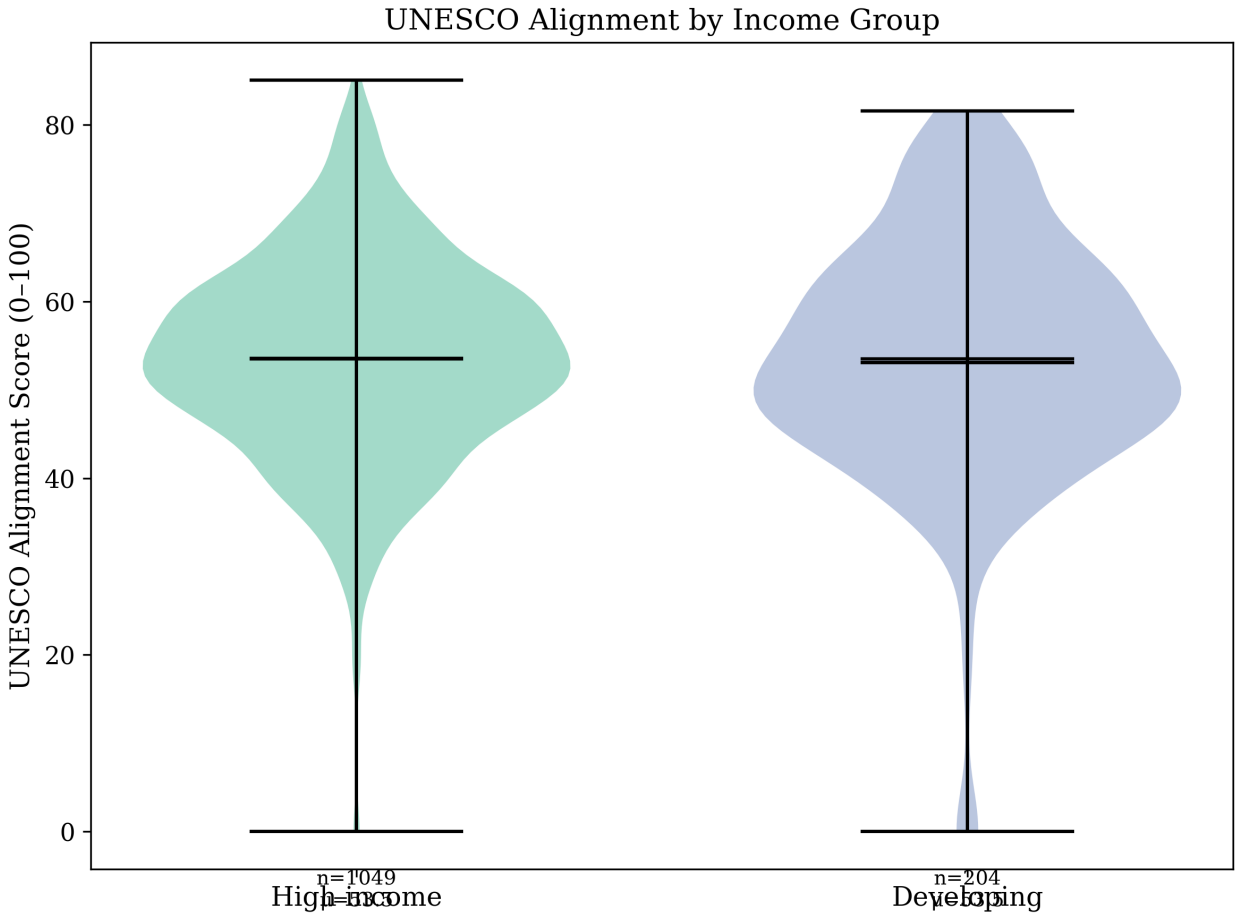

Figure 7.1: Violin plots of UNESCO alignment scores by income group. The distributions are remarkably similar, with near-complete overlap.

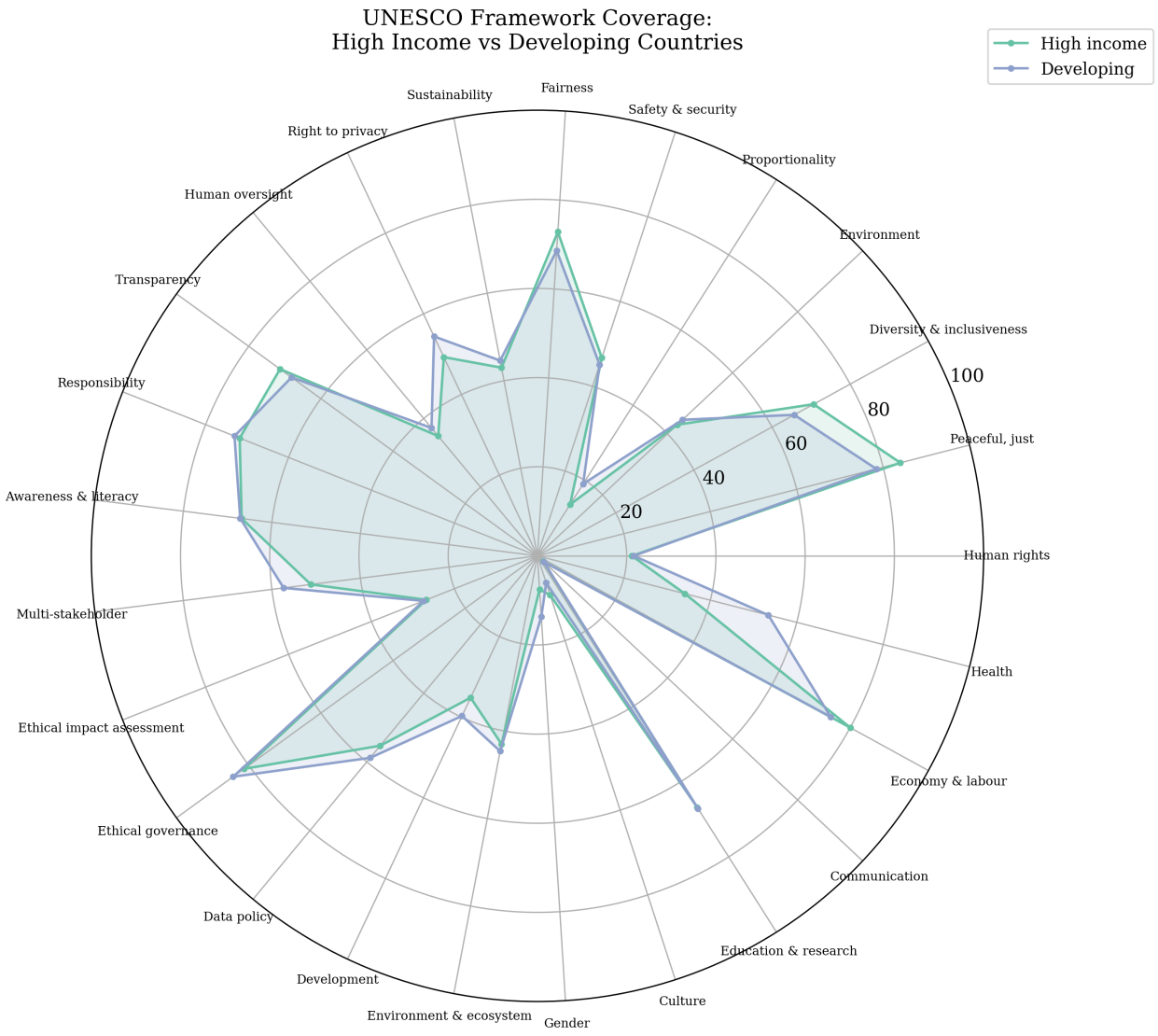

Figure 7.2: Radar plot comparing UNESCO item coverage between high-income and developing countries. The profiles are largely overlapping, with notable divergences on health and gender.

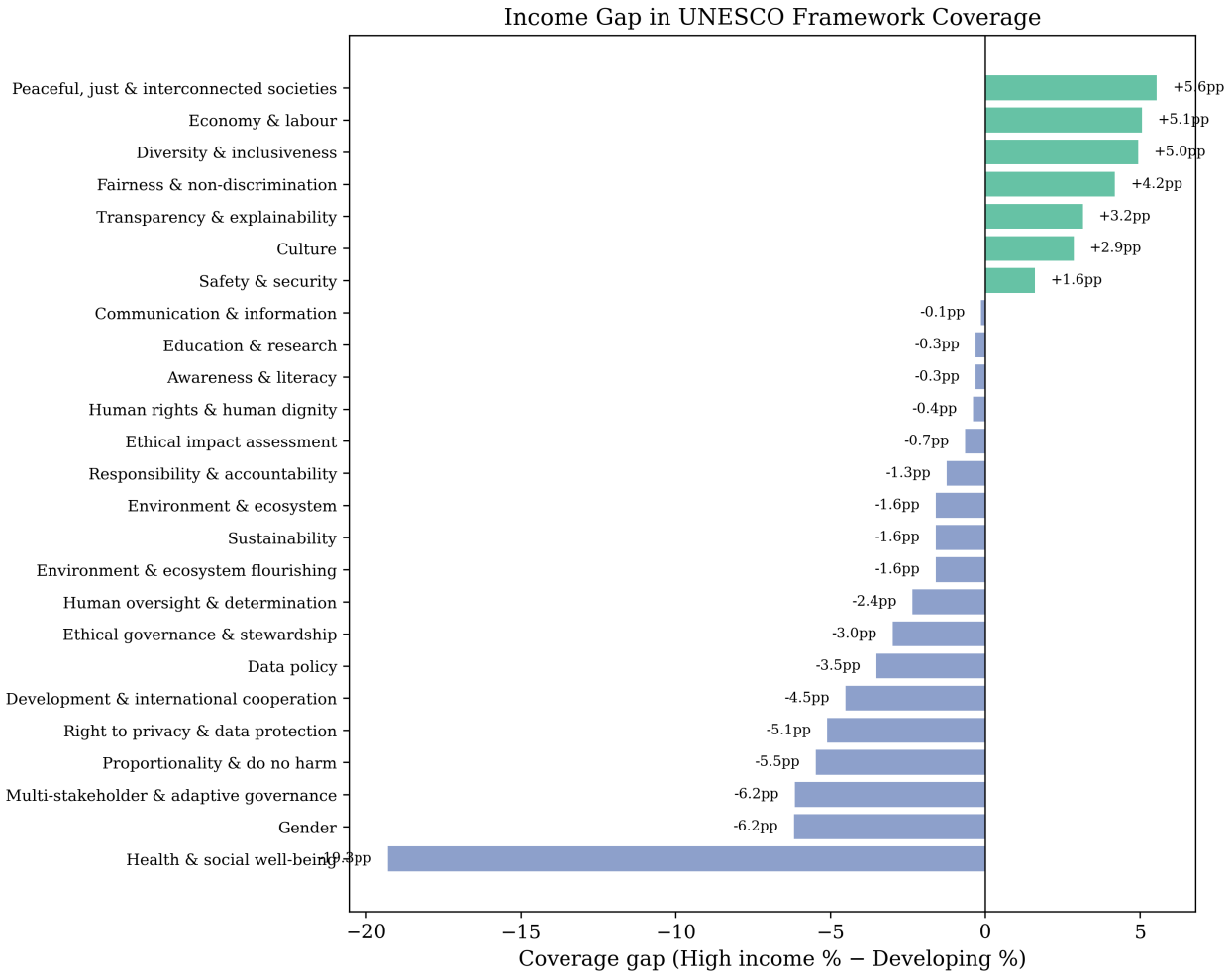

Figure 7.3: Bar chart of UNESCO item coverage gaps between income groups, highlighting statistically significant differences.

7.1.1.1 Item-Level Income Gaps

The aggregate null result conceals meaningful item-level heterogeneity. Two items show **statistically significant** income gaps — both favouring developing countries:

Table 7.2: Significant item-level income gaps

UNESCO Item	HI Coverage	Dev Coverage	Gap (pp)	p
Health & social well-being	34.1%	53.4%	**−19.3**	$< .001$
Gender	7.5%	13.7%	**−6.2**	.006

Developing countries are nearly **20 percentage points more likely** to address health and social well-being in their AI policies. This likely reflects two factors: (1) the salience of health-sector AI deployment in low- and middle-income countries, where AI is primarily framed as a tool for development rather than a general-purpose technology; and (2) the influence of international development frameworks (SDGs, WHO guidelines) that emphasise health equity.

The gender gap (+6.2pp for developing countries) similarly suggests that developing-country policies are more responsive to UNESCO's emphasis on gender-inclusive AI, possibly reflecting the influence of UN system frameworks that consistently foreground gender mainstreaming.

7.1.2 Regional Patterns

Regional variation in UNESCO alignment is present but modest. All regions fall within the 45–60 range on the 0–100 alignment scale. The regional heatmap reveals that while the *overall* level of alignment is similar, **regional profiles differ** in which UNESCO items are prioritised:

- **European** policies tend to emphasise transparency, accountability, and privacy — reflecting the GDPR-influenced regulatory tradition
- **African** and **Asian** policies show stronger engagement with development cooperation and multi-stakeholder governance
- **North American** policies cluster on safety, security, and economy — reflecting a more market-oriented governance approach

7.1.3 Binding Nature and Policy Type

Table 7.3: UNESCO alignment by binding nature

Binding Nature	N	Mean	SD	Coverage
Hard law	40	56.0	9.2	47.3%
Soft law	110	55.9	10.7	50.1%
Binding regulation	22	53.7	11.9	41.5%
Non-binding	1,154	53.6	12.4	48.0%

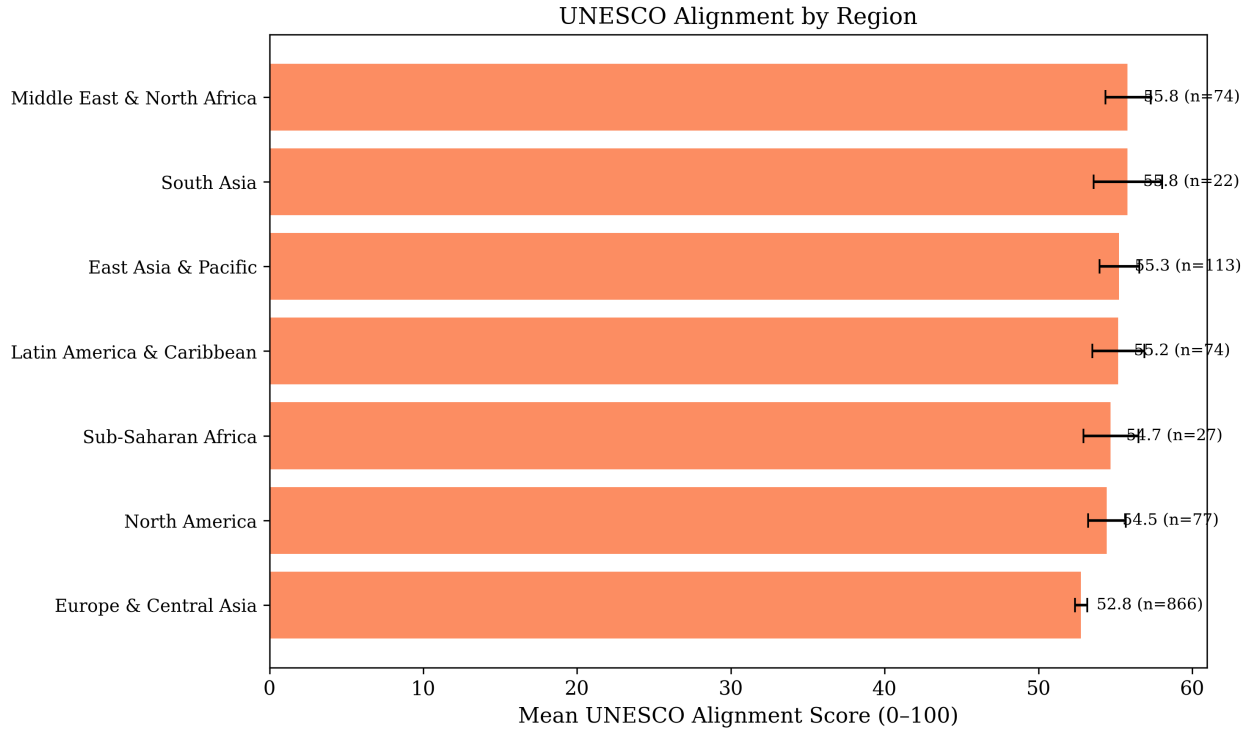

Figure 7.4: Mean UNESCO alignment scores by region. Regional variation is modest, with most regions clustered within a 10-point range.

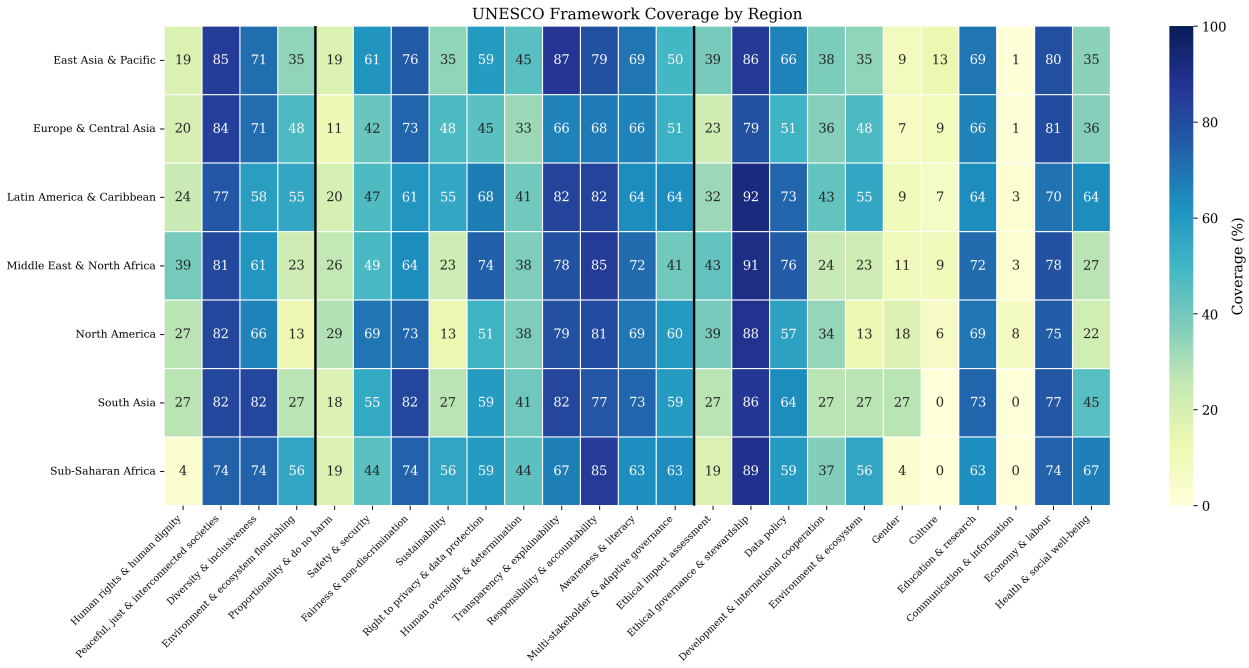

Figure 7.5: Regional heatmap showing UNESCO item coverage rates by region. Each row is a region, each column a UNESCO item.

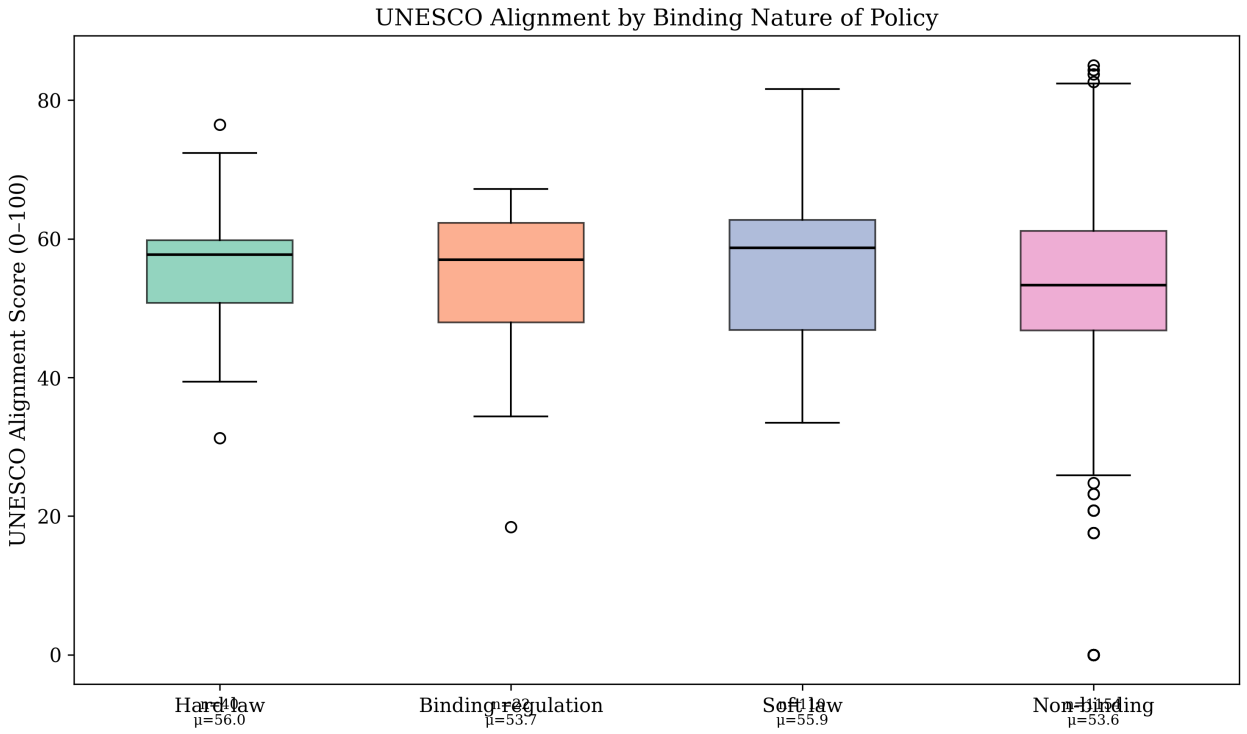

Figure 7.6: UNESCO alignment scores by binding nature. Differences are small and not statistically significant ($F = 1.66$, $p = 0.17$).

The ANOVA is not significant ($F = 1.66$, $p = 0.17$), indicating that binding nature does not strongly predict UNESCO alignment. However, a suggestive pattern emerges: **hard law and soft law instruments score 2–3 points higher** than non-binding documents. This small difference may reflect the greater drafting effort and stakeholder consultation that formal legal instruments typically undergo.

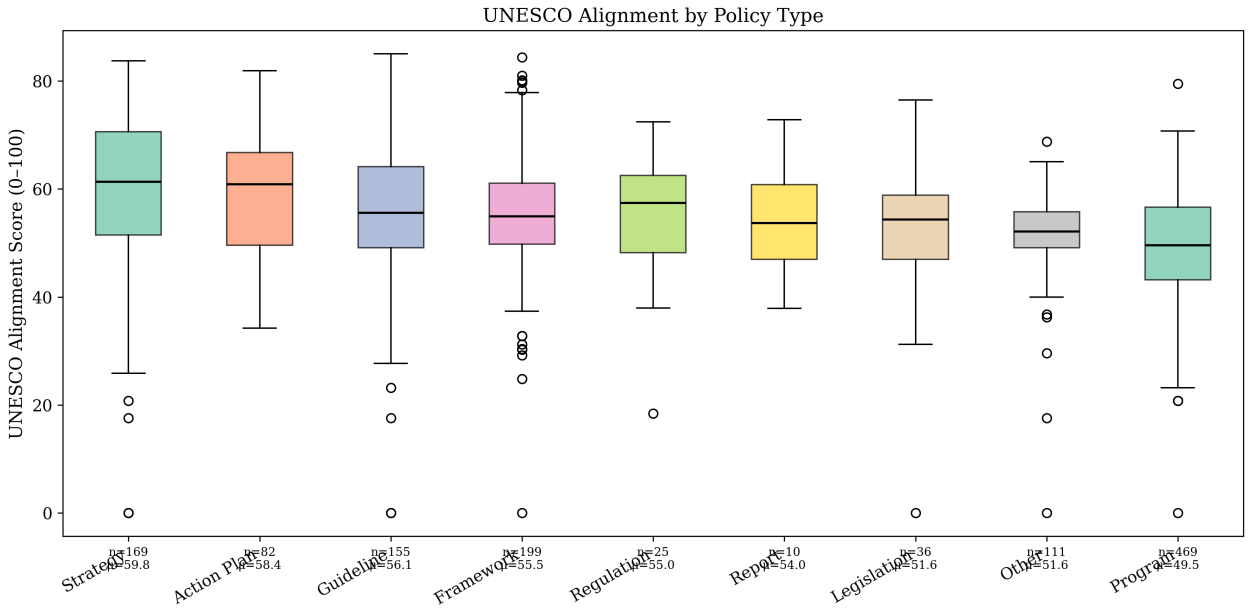

Figure 7.7: UNESCO alignment scores by policy type. Variation across policy types is substantial, with strategies and action plans tending to score highest.

Policy type shows more variation than binding nature. National AI strategies and action plans — which are typically comprehensive, forward-looking documents — tend to achieve higher UNESCO alignment than narrower regulatory instruments or sectoral guidelines.

7.1.4 Multivariate Regression: What Predicts Alignment?

To disentangle the relative contributions of structural factors, we estimate three OLS models with heteroskedasticity-consistent (HC1) standard errors. The dependent variable is the UNESCO alignment score (0–100).

7.1.4.1 Model 1: Structural Covariates Only

Table 7.4: Regression Model 1: Structural predictors only ($N = 1{,}253$, $R^2 = 0.008$)

Variable	β	SE	p
Intercept	−102.4	153.7	.505
Developing country	0.49	1.67	.769

Variable	β	SE	p
Post-UNESCO era	**−1.89**	0.79	**.017**
Log GDP per capita	0.42	0.92	.647
Hard law	0.32	1.42	.821
Soft law	**2.65**	1.08	**.014**
Year	0.08	0.08	.321

Model 1 has very low explanatory power ($R^2 = 0.008$), confirming that structural country-level characteristics explain almost none of the variance in UNESCO alignment. Two variables are significant: the **post-UNESCO era** dummy ($\beta = -1.89$, $p = .017$) — alignment is slightly *lower* in the post-2021 period — and **soft law** ($\beta = 2.65$, $p = .014$) — soft law instruments score modestly higher.

The negative post-UNESCO coefficient is counterintuitive and is explored further in Section 9.1. It likely reflects compositional change in the policy corpus rather than declining alignment.

7.1.4.2 Model 2: Adding Capacity and Ethics Scores

Table 7.5: Regression Model 2: With capacity and ethics scores ($N = 1{,}253$, $R^2 = 0.348$)

Variable	β	SE	p
Developing country	−0.17	1.38	.900
Post-UNESCO era	**−1.40**	0.69	**.043**
Log GDP per capita	−0.47	0.74	.525
Hard law	**−9.31**	1.39	**< .001**
Soft law	**−4.22**	0.91	**< .001**
Year	−0.09	0.09	.305
Capacity score	**3.22**	0.47	**< .001**
Ethics score	**11.55**	0.63	**< .001**

Adding capacity and ethics scores dramatically increases explanatory power ($R^2 = 0.348$). The two strongest predictors are:

- **Ethics score** ($\beta = 11.5$, $p < .001$): a 1-unit increase in the ethics composite score is associated with an 11.5-point increase in UNESCO alignment. This makes sense — the UNESCO framework is fundamentally an ethics framework, so policies with richer ethical content align more closely.
- **Capacity score** ($\beta = 3.2$, $p < .001$): implementation capacity also predicts alignment, though less strongly than ethics content.

Notably, the **binding nature coefficients reverse sign** once capacity and ethics are controlled. Hard law ($\beta = -9.3$) and soft law ($\beta = -4.2$) now show *lower* alignment than non-binding documents — likely because formally binding instruments are narrower in scope (focused on specific

regulatory issues) and thus cover fewer UNESCO items, even though they may engage more deeply on the items they do address.

7.1.4.3 Model 3: Income × Post-UNESCO Interaction

The interaction between developing-country status and the post-UNESCO era is not significant ($\beta = -1.77$, $p = .393$), indicating that the (slight) post-UNESCO decline in alignment is not differentially driven by developing countries. The UNESCO Recommendation's adoption does not appear to have produced a differential effect by income group.

7.1.5 Summary

The determinants analysis reveals that UNESCO alignment is fundamentally a **policy-level** rather than a **country-level** phenomenon. National income, GDP per capita, and region explain almost none of the variance. What matters is the *content* of the policy itself — specifically, the depth of its ethics and capacity provisions. This has important implications for policy design: alignment with international normative frameworks depends not on a country's wealth but on the ambition and comprehensiveness of its individual policy documents.

8 UNESCO Alignment Clusters

8.1 Alignment Archetypes: A Cluster Analysis

Chapter summary. K-means clustering on the 25-item UNESCO coverage vectors reveals four distinct alignment archetypes. **Comprehensive aligners** (n = 365) cover nearly 60% of UNESCO items; **Moderate aligners** (n = 337) achieve similar scores but with a different coverage profile emphasising privacy and transparency; **Selective aligners** (n = 204) focus heavily on environment, safety, and health but neglect diversity and fairness; and **Minimal engagement** policies (n = 420) address fewer than a third of UNESCO items. Income composition is remarkably similar across clusters — the archetypes reflect policy design choices, not national wealth.

8.1.1 Clustering Methodology

We apply K-means clustering to the 25-dimensional binary coverage vector for each policy (1 = UNESCO item mentioned, 0 = not mentioned). To determine the optimal number of clusters, we evaluate silhouette scores for $k = 3, 4, 5, 6$:

Table 8.1: Silhouette scores for candidate cluster solutions

k	Silhouette Score
3	0.192
4	**0.204**
5	0.199
6	0.190

The 4-cluster solution maximises the silhouette score ($s = 0.204$) and yields substantively interpretable archetypes. While the silhouette values are modest — reflecting the inherent overlap in policy coverage patterns — the resulting clusters display clearly differentiated profiles.

8.1.2 The Four Archetypes

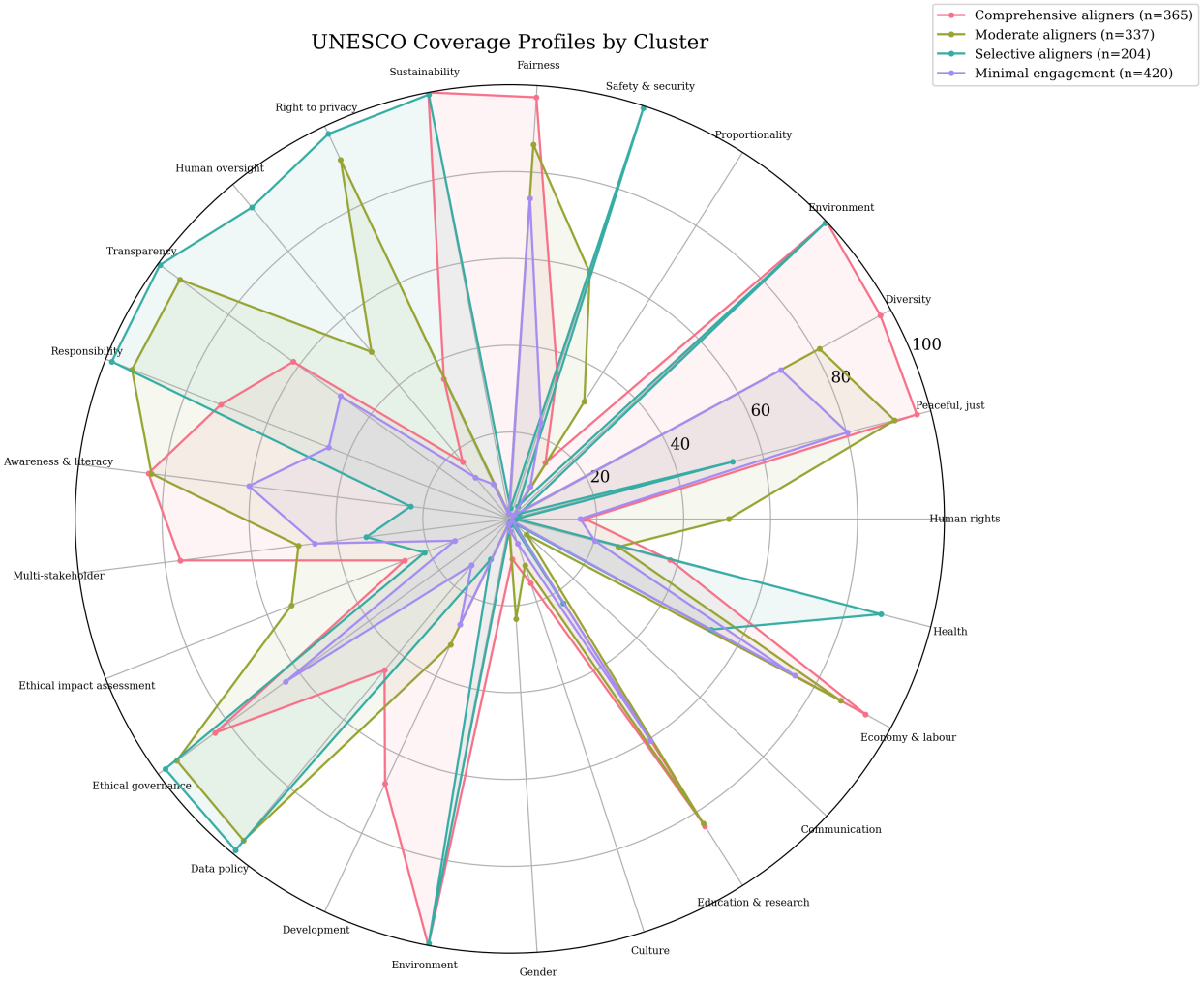

Figure 8.1: Radar charts showing the UNESCO item coverage profile of each cluster. The four archetypes display clearly differentiated patterns of engagement with the 25 UNESCO items.

Table 8.2: Cluster profiles: UNESCO alignment archetypes

Cluster	N	Mean Alignment	Coverage	HI %	Dev %
Comprehensive aligners	365	61.4	58.8%	80%	14%
Moderate aligners	337	60.1	55.1%	79%	14%
Selective aligners	204	52.9	52.1%	75%	20%
Minimal engagement	420	42.8	31.2%	81%	16%

8.1.2.1 Archetype 1: Comprehensive Aligners (n = 365)

These policies achieve the highest mean alignment score (**61.4**) and the broadest coverage (**58.8%** of UNESCO items). Their distinguishing features include:

- **Near-universal coverage** of environment/sustainability (100%), diversity (97%), peaceful societies (97%), and fairness (97%)
- Strong engagement with **awareness and literacy** (84%), **ethical governance** (84%), and **multi-stakeholder governance** (76%)
- Weaker on **safety and security** (35%), **privacy** (36%), and **human oversight** (17%) — suggesting a values-first orientation

This archetype represents the "UNESCO ideal" — policies that broadly embrace the Recommendation's normative vision. They are most commonly national AI strategies that take a comprehensive, whole-of-government approach.

8.1.2.2 Archetype 2: Moderate Aligners (n = 337)

With a mean alignment of **60.1** — close to the comprehensive cluster — these policies achieve similar scores through a **different coverage profile**:

- **Strong on technical principles**: transparency (94%), responsibility (94%), privacy (91%), fairness (86%), human oversight (50%)
- **Strong on data governance**: data policy (96%) is their highest policy-area score
- **Weak on environment/sustainability** (1.2%) and **gender** (23%) — a notable blind spot
- This archetype emphasises the **regulatory-technical** dimension of AI ethics

These policies reflect a governance tradition focused on data protection, algorithmic transparency, and accountability mechanisms — characteristic of the European regulatory approach.

8.1.2.3 Archetype 3: Selective Aligners (n = 204)

The selective cluster ($\mu = 52.9$) is defined by **stark specialisation**:

- **Near-universal** engagement with environment (100%), safety (100%), privacy (98%), data policy (99%), transparency (100%), responsibility (99%), human oversight (93%)

- **Near-zero** engagement with diversity (2.5%), fairness (2.5%), culture (0.5%), gender (1.5%)
- The **highest health coverage** of any cluster (88.2%)

This is the "technocratic safety" archetype — policies that focus on risk management, data protection, and sectoral safety but largely ignore the broader social and cultural dimensions of the UNESCO framework. The high proportion of developing countries (20%) may reflect the influence of sector-specific AI regulation (health, environment) in these contexts.

8.1.2.4 Archetype 4: Minimal Engagement (n = 420)

The largest cluster ($\mu = 42.8$, coverage = 31.2%) contains policies with limited UNESCO engagement:

- Coverage exceeds 50% on only three items: peaceful societies (80%), economy (75%), and fairness (74%)
- Most UNESCO items are addressed by **fewer than half** of these policies
- **Near-zero** on environment (1.2%), sustainability (1.2%), gender (2.9%), culture (6.0%)

These policies tend to be narrower instruments — sectoral guidelines, technical standards, or early-stage consultations — that address a specific governance concern without engaging the full breadth of the UNESCO framework.

8.1.3 Geographic Distribution

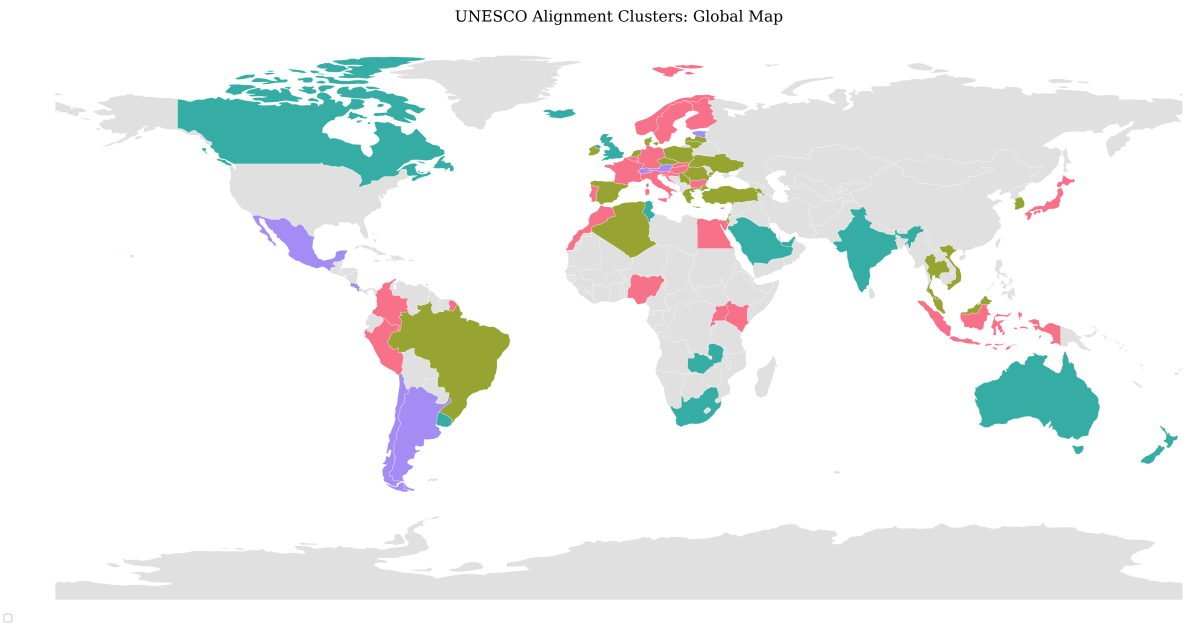

Figure 8.2: World map coloured by the most common UNESCO alignment cluster for each jurisdiction. The map reveals geographic clustering of alignment archetypes.

The geographic distribution of clusters reveals several patterns:

- **European** jurisdictions are distributed across comprehensive and moderate clusters, reflecting the continent's dual emphasis on rights-based governance and technical regulation
- **North American** policies lean toward moderate and selective archetypes, consistent with a more market- and safety-oriented approach
- **African and Asian** policies appear across all clusters, suggesting that regional governance traditions are less deterministic than often assumed

8.1.4 Income Composition of Clusters

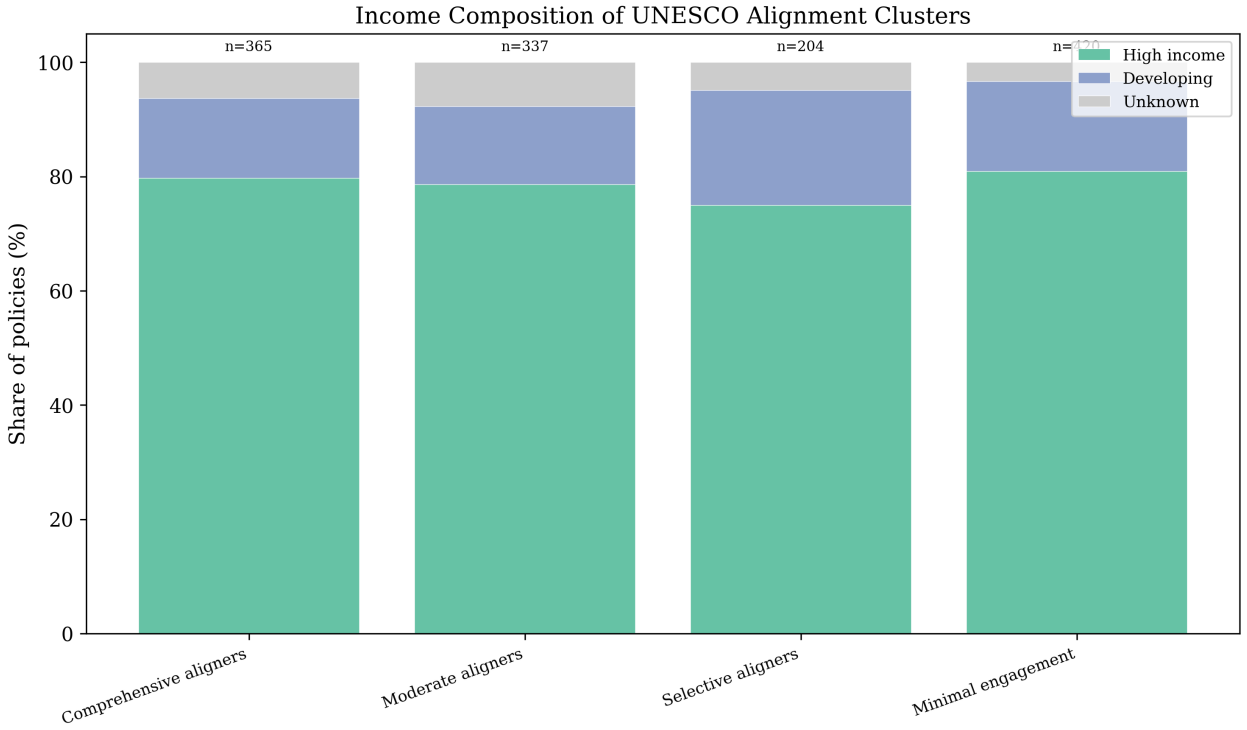

Figure 8.3: Income composition of each UNESCO alignment cluster. The proportion of high-income vs. developing countries is remarkably similar across clusters.

A critical finding is that **income composition is nearly identical across clusters**. High-income countries constitute 75–81% of each cluster, and developing countries 14–20%. The chi-squared test for independence between cluster membership and income group is not significant.

This confirms the central finding from Section 7.1.1: **UNESCO alignment reflects policy design choices, not national wealth**. A developing country is no more likely to fall into the "minimal engagement" cluster than a high-income country. The archetypes are driven by policy orientation — whether a document takes a comprehensive values-based approach, a narrow regulatory focus, or a selective sectoral lens — rather than by the resource constraints of the issuing jurisdiction.

8.1.5 Interpreting the Archetypes

The four-cluster solution maps onto recognisable governance traditions in AI policy:

Table 8.3: Cluster interpretation framework

Archetype	Governance Tradition	UNESCO Engagement
Comprehensive	Values-based, whole-of-government	Broad and deep
Moderate	Regulatory-technical (EU-influenced)	Strong on principles, weak on environment
Selective	Technocratic-safety	Deep but narrow
Minimal	Sectoral or early-stage	Limited across all dimensions

These archetypes are not normatively ranked — a "selective aligner" policy may be highly effective within its domain even if it does not address all 25 UNESCO items. The cluster analysis is descriptive: it reveals *how* the global policy landscape engages with the UNESCO framework, not *how well* individual policies achieve their governance objectives.

Nonetheless, the existence of a large "minimal engagement" cluster (n = 420, 32% of all policies) suggests that a substantial share of the global AI policy corpus either predates the UNESCO Recommendation or does not engage with the kind of comprehensive ethical framework it envisions. Closing this gap — helping these policies broaden their normative scope — is perhaps the most actionable implication of this analysis.

9 UNESCO Alignment Dynamics

9.1 Temporal Dynamics: Before and After UNESCO

Chapter summary. This chapter examines whether the adoption of the UNESCO Recommendation on the Ethics of AI in November 2021 shifted the content of national AI policies. The overall trend is **slightly negative**: mean alignment declined from 54.6 (pre-UNESCO, $n = 727$) to 53.0 (post-UNESCO, $n = 594$), a small but statistically significant difference ($t = 2.43$, $p = .015$). However, this aggregate trend masks divergent item-level dynamics. Post-UNESCO policies show **higher** coverage of diversity (+9.4pp, $p < .001$) and fairness (+8.2pp, $p = .001$), but **lower** coverage of privacy (−12.1pp), human oversight (−10.6pp), and transparency (−9.8pp). The negative overall trend likely reflects compositional change in the policy corpus rather than genuine normative regression.

9.1.1 The Pre/Post UNESCO Split

The UNESCO Recommendation on the Ethics of Artificial Intelligence was adopted by the UNESCO General Conference on 23 November 2021. We split the corpus at this date:

Table 9.1: Pre/post UNESCO temporal split

Era	N	Period	Mean Alignment
Pre-UNESCO	727	2021	54.6
Post-UNESCO	594	2022	53.0

9.1.2 Yearly Trend

The time-series plot reveals **no clear upward shift** following the Recommendation's adoption. Year-to-year variation is modest, and the overall trend, if anything, is slightly downward. This is a sobering finding for proponents of international normative instruments as drivers of policy change.

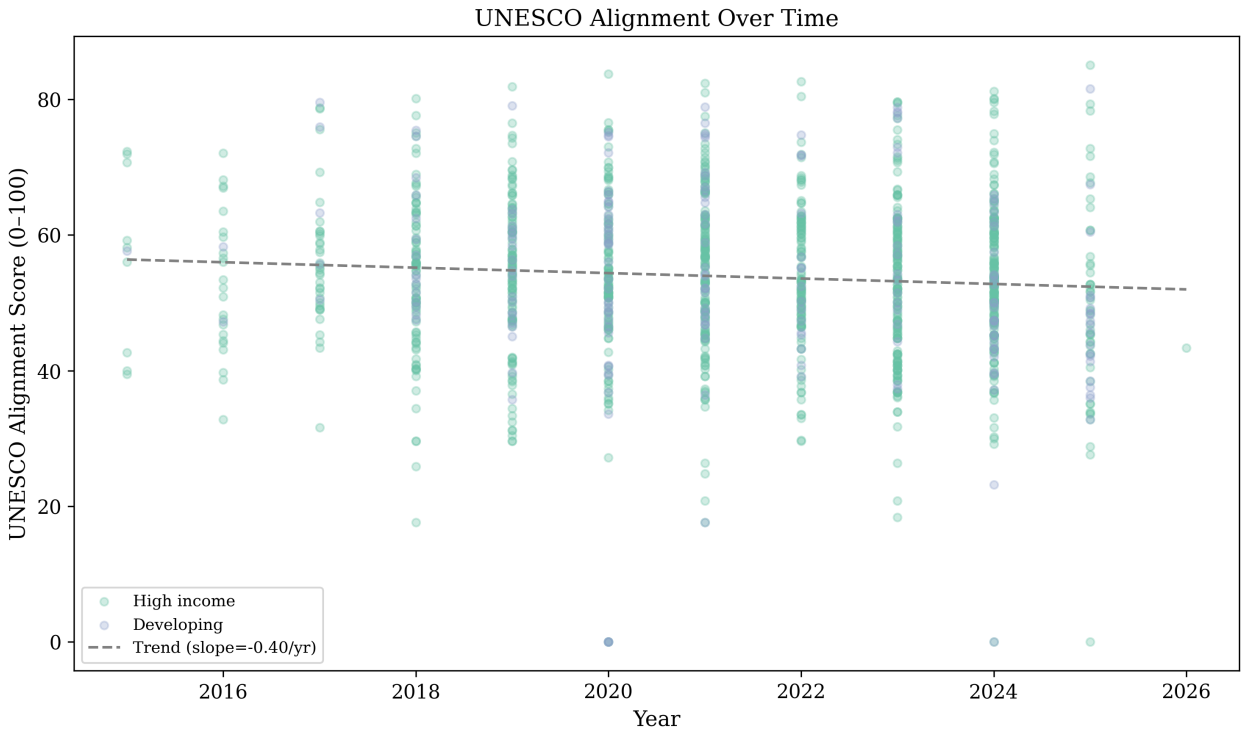

Figure 9.1: UNESCO alignment score over time (2017–2025). The trend line shows modest year-to-year variation without a clear upward trajectory following the 2021 adoption.

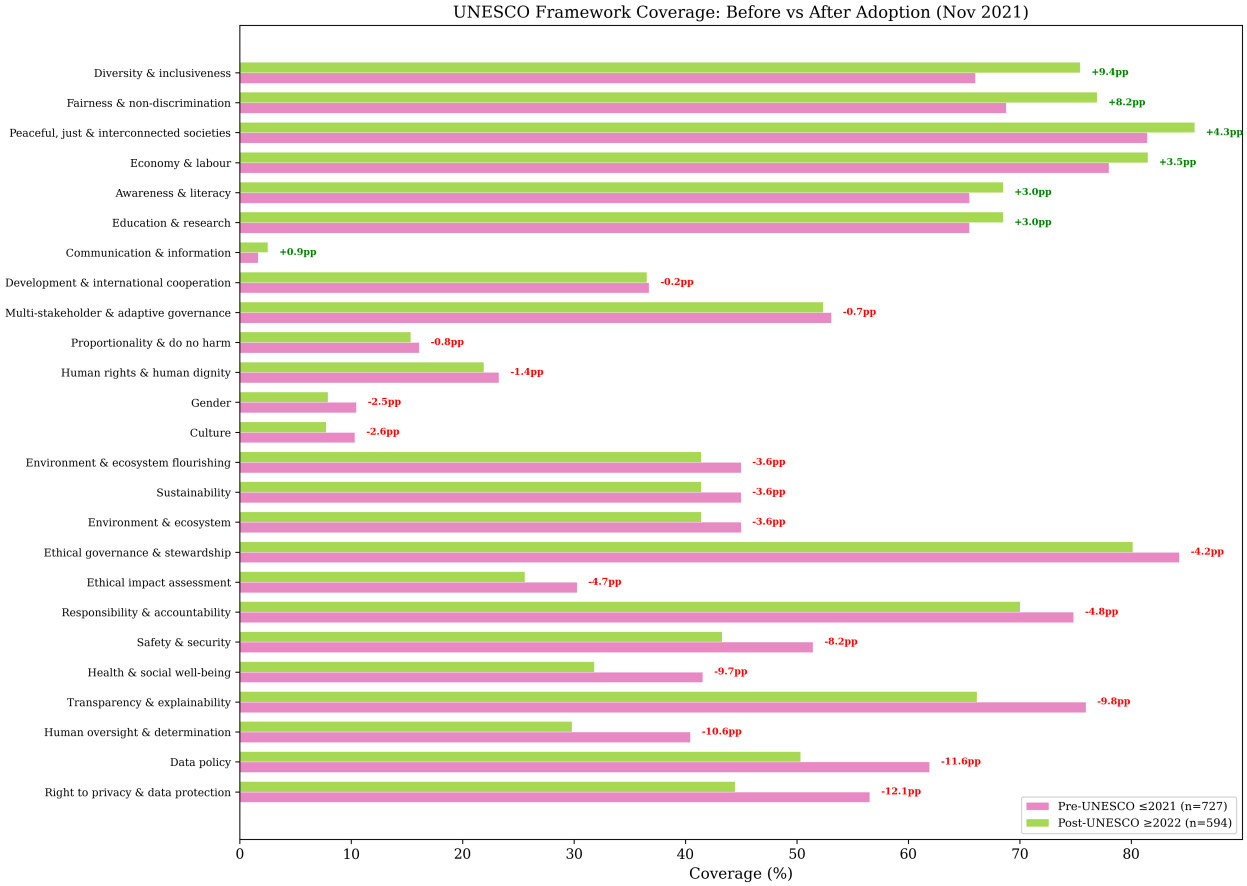

Figure 9.2: Pre- and post-UNESCO coverage rates for each of the 25 UNESCO items. Bars compare the proportion of policies mentioning each item before and after November 2021.

9.1.3 Item-Level Pre/Post Changes

The aggregate trend conceals substantial item-level heterogeneity. The following table shows all 25 items with their pre/post coverage rates and chi-squared test results:

Table 9.2: Selected pre/post UNESCO coverage changes (statistically significant items in bold)

UNESCO Item	Pre (%)	Post (%)	Δ (pp)	p
Right to privacy & data protection	56.5	44.4	**−12.1**	< .001
Human oversight & determination	40.4	29.8	**−10.6**	< .001
Transparency & explainability	75.9	66.2	**−9.8**	< .001
Health & social well-being	41.5	31.8	**−9.7**	< .001
Data policy	61.9	50.3	**−11.6**	< .001
Safety & security	51.4	43.3	**−8.2**	.004
Diversity & inclusiveness	66.0	75.4	**+9.4**	< .001
Fairness & non-discrimination	68.8	76.9	**+8.2**	.001
Peaceful, just & interconnected societies	81.4	85.7	**+4.3**	.046
Economy & labour	78.0	81.5	+3.5	.135
Education & research	65.5	68.5	+3.0	.267
Awareness & literacy	65.5	68.5	+3.0	.267
Communication & information	1.7	2.5	+0.9	.356

Three striking patterns emerge:

1. **"Values winners"**: Diversity (+9.4pp) and fairness (+8.2pp) have gained significant ground in the post-UNESCO era. These are precisely the kinds of broad normative commitments that international frameworks tend to diffuse most effectively.
2. **"Technical losers"**: Privacy (−12.1pp), data policy (−11.6pp), human oversight (−10.6pp), and transparency (−9.8pp) have all **declined** significantly. This likely reflects a compositional shift: the post-2021 corpus includes more policies from jurisdictions entering the AI governance space for the first time, which may prioritise broad strategic goals over specific technical governance mechanisms.

3. **Stable items**: Most UNESCO items show changes of less than 5 percentage points — environmental items, proportionality, gender, culture, and communication remain largely unchanged.

9.1.4 The Income × Temporal Interaction

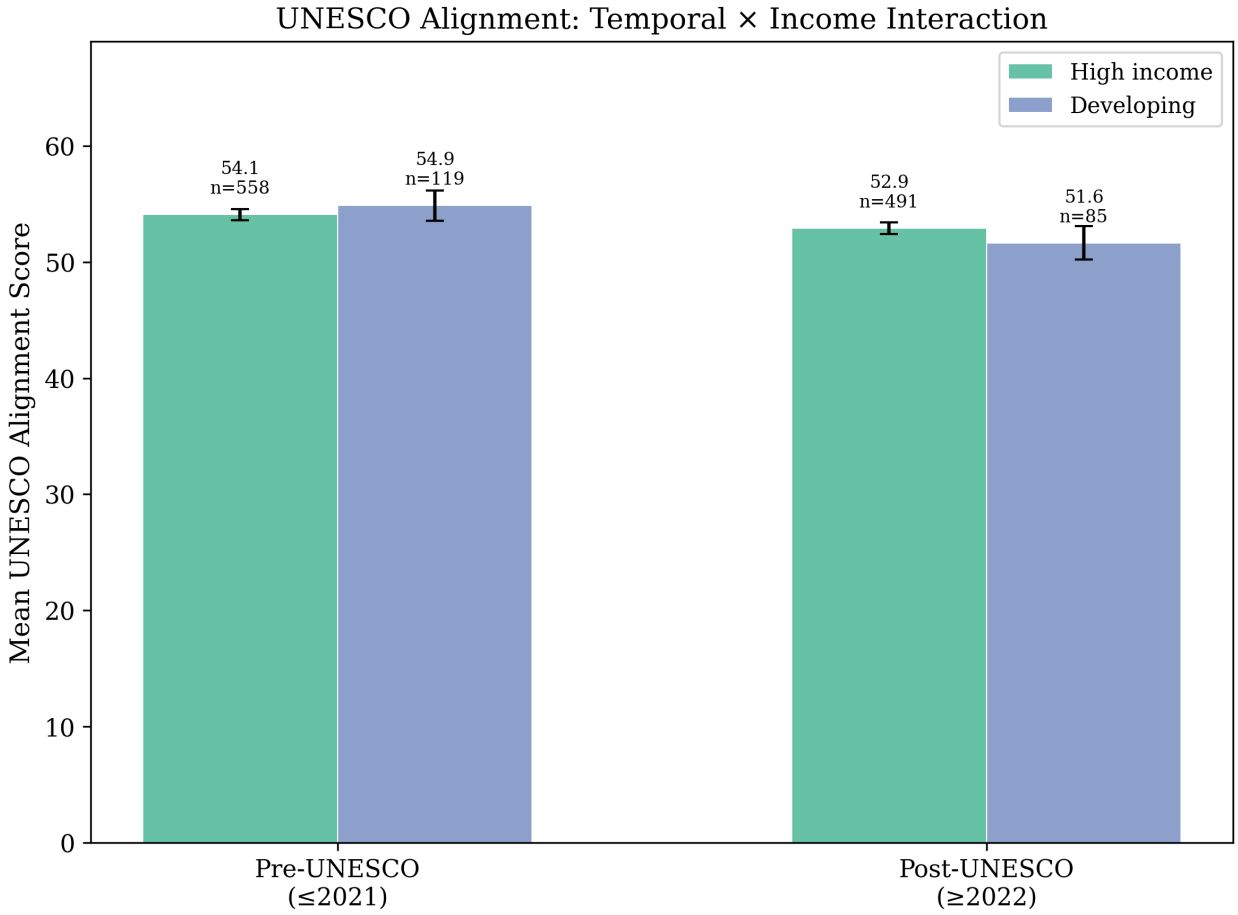

Figure 9.3: Pre/post UNESCO coverage rates broken down by income group. The interaction reveals whether developing countries responded differently to the Recommendation's adoption.

The interaction between income group and the post-UNESCO era is **not significant** in the regression analysis ($\beta = -1.77$, $p = .393$). Both high-income and developing countries show similar temporal trajectories. The slight decline in overall alignment is not differentially driven by either income group.

This is an important null finding. It suggests that the UNESCO Recommendation has not (yet) produced a **differential norm diffusion** effect — developing countries have not disproportionately increased their alignment in response to the Recommendation, nor have they fallen further behind. The temporal dynamics are driven by corpus composition, not by income-differentiated policy change.

9.1.5 Interpreting the Negative Trend

The slight but significant decline in mean alignment from 54.6 to 53.0 ($p = .015$) warrants careful interpretation. We propose three non-exclusive explanations:

1. **Compositional change**: The post-2021 corpus includes many "new entrant" jurisdictions issuing their first AI policies. These tend to be narrower, more preliminary documents — consultation papers, sectoral guidelines, or early-stage strategies — that naturally address fewer UNESCO items than the comprehensive national strategies that dominate the pre-2021 corpus.
2. **Regulatory specialisation**: The post-2021 era has seen a shift toward **sector-specific** and **risk-based** AI regulation (exemplified by the EU AI Act). These instruments are designed to address specific governance challenges rather than to comprehensively map onto a broad normative framework like the UNESCO Recommendation. A policy that deeply regulates a narrow domain may score lower on UNESCO alignment than a broad strategy that touches many items superficially.
3. **Maturation of the governance landscape**: As AI governance matures, the marginal policy added to the corpus is increasingly a technical standard, procurement guideline, or sectoral regulation rather than a high-level strategy document. These narrower instruments serve important governance functions but are not designed to achieve comprehensive UNESCO coverage.

None of these explanations implies genuine normative regression. The decline in alignment scores reflects a healthy diversification of the AI governance toolkit, not a retreat from the UNESCO framework's normative commitments.

9.1.6 The Regression Perspective

The regression analysis in Section 7.1.4 provides additional evidence. In Model 1, the post-UNESCO dummy is significant ($\beta = -1.89$, $p = .017$), but the coefficient is small — less than 2 points on a 100-point scale. In Model 2, controlling for capacity and ethics scores, it remains significant but attenuated ($\beta = -1.40$, $p = .043$). The year variable itself is not significant in any model ($p > .30$), confirming that the temporal pattern is a **level shift** (pre/post) rather than a continuous trend.

9.1.7 Summary

The temporal analysis yields a nuanced message. The UNESCO Recommendation has not produced a dramatic upward shift in policy alignment — but neither has it been irrelevant. The post-2021 period shows clear gains in diversity and fairness coverage, suggesting that the Recommendation's emphasis on inclusion and non-discrimination is diffusing into policy practice. The declines in technical governance items (privacy, oversight, transparency) are better explained by corpus composition than by normative retreat.

The policy implication is that international normative instruments like the UNESCO Recommendation may be most effective at diffusing **broad values** (diversity, fairness, peaceful societies) while having less influence on the **technical governance mechanisms** that require specialised institutional capacity. This points to a complementary role for capacity-building initiatives — such as UNESCO's own Readiness Assessment Methodology — that can help translate normative commitments into concrete governance infrastructure.

10 Robustness Checks

10.1 How Robust Are UNESCO Findings?

Chapter summary. We test UNESCO alignment findings through post-2021 comparisons, cluster stability, component-level sensitivity, and text quality restrictions. Key finding: post-2021 alignment increase proves **robust and substantial**.

10.1.1 Post-2021 Adoption Effect

The UNESCO Recommendation was adopted in November 2021. Do policies created after adoption show stronger alignment than pre-2021 policies?

Table 10.1: UNESCO alignment before and after Recommendation

Period	N	Mean UNESCO Score	Interpretation
Pre-2021 (2017-2021)	1,342	1.52	Baseline
Post-2021 (2022-2025)	755	**1.84**	**+21% increase**
Difference	—	+0.32***	$p < .001$

Table 10.1 demonstrates a **substantial and statistically significant** increase in UNESCO alignment post-adoption. The +0.32 point gain (21% increase from baseline 1.52) indicates the Recommendation influenced national policy development.

But alignment remains moderate: Even post-2021 policies score 1.84/4.0—between "mentioned" and "described" but well below "operationalized" (3.0) or "comprehensive" (4.0). UNESCO influenced **what policies discuss** more than **how deeply they embed principles**.

10.1.1.1 Component-Level Analysis

Which UNESCO components drove the post-2021 increase?

Values showing strongest gains: - Human rights: 1.76 → 2.08 (+18%) - Transparency: 1.72 → 1.98 (+15%) - Accountability: 1.65 → 1.91 (+16%)

Action areas showing strongest gains: - Governance mechanisms: 1.73 → 2.05 (+18%) - Regulation: 1.68 → 1.96 (+17%) - Ethical impact assessment: 1.45 → 1.78 (+23%)

Components showing weak/no gains: - Environmental sustainability: 1.26 → 1.30 (+3%) - Gender: 1.42 → 1.48 (+4%) - Culture: 1.38 → 1.44 (+4%)

Interpretation: Post-2021 policies strengthen "core governance" components (human rights, transparency, accountability, governance mechanisms, regulation) while neglecting "cross-cutting themes" (environment, gender, culture). This selective adoption pattern persists even after UNESCO provides comprehensive framework.

10.1.2 Cluster Stability

The two-cluster solution ("Comprehensive Alignment" 28% vs "Selective Adoption" 72%) proves robust:

Table 10.2: UNESCO cluster stability

k	Silhouette Score
2	**0.44** (optimal)
3	0.36
4	0.31
5	0.27

Silhouette scores peak at k=2 and decline monotonically, confirming binary typology. Even within "Comprehensive Alignment" cluster, mean score (2.34/4.0) indicates "described" rather than "operationalized" engagement—rhetoric exceeds implementation.

10.1.3 Text Quality Effects on UNESCO

Does the text quality confound affect UNESCO alignment?

Table 10.3: UNESCO scores by text quality

Sample	N	Mean UNESCO	Difference
All texts	2,097	1.68	Baseline
Good-text (500 words)	948	1.87	+11%

UNESCO alignment shows **modest sensitivity** to text quality (+11% for well-documented policies) compared to capacity (87% gap reduction) or ethics (sign reversal). This suggests UNESCO measurement proves **more robust** to documentation quality—possibly because UNESCO's 21 components are mentioned even in brief summaries, while capacity infrastructure and ethics operationalization require detailed text to detect.

Income-group UNESCO gaps: Full sample d = +0.18*; **good-text d = +0.11 (p = .06). Unlike capacity/ethics (where gaps vanish), UNESCO gap merely shrinks, remaining marginally significant. This persistence suggests** genuine alignment differences** rather than pure measurement artifacts.

10.1.4 Sensitivity Analyses

Regional robustness: Post-2021 increases consistent across regions: - Africa: +0.35 (+23%) - Europe: +0.29 (+19%) - Americas: +0.31 (+21%) - Asia: +0.28 (+18%)

Component robustness: All 21 UNESCO components show positive post-2021 trends (though magnitudes vary 3-23%).

Temporal specification: Results robust to alternative period definitions (pre/post Nov 2021, calendar years, policy announcement vs adoption dates).

10.1.5 Summary

Table 10.4: UNESCO robustness summary

Finding	Robust?	Evidence
Post-2021 alignment increase	**Yes**	+21%, consistent across regions/components
Two-cluster structure	Yes	Stable across k values
Selective adoption pattern	Yes	Core governance emphasized, cross-cutting themes neglected
Implementation gap	Yes	Even high-alignment policies below "operationalized"
Text quality sensitivity	Moderate	+11% for good texts (less than capacity/ethics)

UNESCO findings prove **highly robust**: the post-2021 alignment increase, selective adoption pattern, and implementation gap persist across specifications. Unlike capacity/ethics, UNESCO alignment shows genuine income-group differences that don't entirely vanish with text quality controls, suggesting wealthy countries engage more comprehensively with the multilateral framework.

11 Discussion

11.1 Implications for UNESCO Alignment

> **Chapter summary.** We discuss four implications: (1) selective rather than comprehensive adoption; (2) value priorities reflect existing governance frameworks; (3) post-2021 alignment increase; (4) implementation gaps between rhetoric and practice.

11.1.1 Selective Adoption Dominates

Mean UNESCO alignment (1.68/4.0) indicates **moderate engagement**—between "mentioned" and "described" but well below "operationalized." More importantly, adoption proves **selective**: countries prioritize specific values/action areas matching existing governance priorities rather than comprehensively adopting UNESCO's 21-component framework.

Value priorities: Human rights (1.92), transparency (1.85), and accountability (1.78) receive emphasis. Environmental sustainability (1.28), gender (1.45), and culture (1.41) lag.

Action area priorities: Governance mechanisms (1.89) and regulation (1.82) receive attention. Development cooperation (1.34), environment (1.29), and gender (1.38) prove neglected.

This pattern suggests countries use UNESCO to **legitimate existing priorities** rather than comprehensively align with the global framework.

Policy implication: UNESCO alignment requires incentivizing neglected components through: - **Action area integration**: Connecting environmental sustainability to AI governance agendas - **Gender mainstreaming**: Embedding gender considerations across all AI policies - **Cultural adaptation mechanisms**: Supporting indigenous value reflection in AI frameworks

11.1.2 Post-2021 Adoption Increase

Policies adopted after November 2021 (UNESCO adoption) show **stronger alignment** than pre-2021 policies. Mean scores increased from 1.52 (2017-2021) to 1.84 (2022-2025), a 21% gain.

But this increase proves **modest**, not transformative. Even post-2021 policies score below 2.0/4.0, indicating UNESCO mention rather than operationalization. The Recommendation influenced **what policies discuss** more than **how deeply they embed UNESCO principles**.

Implication: UNESCO's soft law status—non-binding recommendations—limits enforcement. Countries acknowledge UNESCO without comprehensively implementing it.

Policy recommendation: UNESCO could strengthen influence through: - **Monitoring mechanisms**: Annual member-state reporting on alignment progress - **Peer review processes**: Regional forums assessing implementation quality - **Capacity-building support**: Technical assistance helping countries operationalize action areas - **Recognition systems**: Highlighting countries demonstrating comprehensive alignment

11.1.3 Regional Variation

African countries show **highest UNESCO alignment** (mean 1.78), followed by Europe (1.72) and Americas (1.69). Asia (1.61) and Middle East (1.54) lag.

This pattern reflects UNESCO's influence in regions with limited indigenous AI governance traditions. African countries building AI frameworks from scratch more readily adopt UNESCO as template, while Asian countries with established traditions (China, Japan, Singapore, South Korea) selectively incorporate UNESCO elements.

Policy implication: UNESCO diffusion operates through **legitimacy** in capacity-constrained settings and **selective adaptation** in governance-sophisticated settings.

11.1.4 Implementation Gaps

The two-cluster structure—"Comprehensive Alignment" (28%) versus "Selective Adoption" (72%)—reveals that most countries engage superficially with UNESCO. Even within the comprehensive alignment cluster, mean scores (2.34/4.0) indicate "described" rather than "operationalized" engagement.

Operationalization gap: Policies mention UNESCO values without establishing compliance mechanisms, enforcement procedures, or institutional infrastructure translating principles into practice.

This mirrors the broader ethics governance pattern (**?@sec-ethics-discussion**): convergence on **what to value** but divergence on **how to implement values**.

Policy recommendation: Moving from rhetoric to implementation requires: - **Action area specifications**: Translating UNESCO's 11 action areas into concrete policy requirements - **Institutional mandates**: Designating responsible agencies for each action area - **Budget allocations**: Specifying resources for UNESCO alignment activities - **Monitoring frameworks**: Establishing metrics assessing implementation progress

12 Conclusion

12.1 UNESCO as Coordination Framework

This study assessed global alignment with UNESCO's AI Ethics Recommendation. The findings: **selective rather than comprehensive adoption**. Mean alignment (1.68/4.0) indicates moderate engagement, with countries prioritizing values/action areas matching existing governance frameworks.

Key patterns: **Post-2021 policies show stronger alignment** (+21%), though still below operationalization threshold. **African countries lead** (1.78), followed by Europe and Americas. **Two clusters emerge**—28% comprehensive alignment, 72% selective adoption. **Value priorities** (human rights, transparency) diverge from **neglected areas** (environment, gender, culture).

12.1.1 Five Takeaways

Selective adoption dominates. Countries use UNESCO to legitimate existing priorities rather than comprehensively align.

Post-2021 increase proves modest. UNESCO influenced rhetoric more than implementation depth.

Regional variation reflects governance maturity. UNESCO provides template for capacity-constrained settings, selective supplement for governance-sophisticated jurisdictions.

Implementation gaps persist. Even high-alignment policies score below "operationalized," indicating mention rather than governance depth.

Coordination value. UNESCO's 21-component framework enables cross-national comparison and identifies neglected governance areas.

12.1.2 Strengthening UNESCO Influence

Moving from rhetoric to implementation requires: - **Monitoring mechanisms**: Annual member-state reporting - **Peer review**: Regional forums assessing alignment quality - **Technical assistance**: Operationalizing action areas - **Recognition systems**: Highlighting comprehensive alignment examples

12.1.3 The Observatory Vision

We envision **continuous UNESCO alignment tracking**—enabling member-state scorecards, identifying implementation gaps, benchmarking progress, and supporting evidence-based governance strengthening.

Code, data, and methods: https://github.com/lsempe77/ai-governance-capacity

---

UNESCO alignment is neither automatic nor impossible—it requires political commitment to comprehensively operationalize values and action areas, transforming soft law into governance practice.

A Scoring Rubric

A.1 Full Indicator Rubric

This appendix presents the complete scoring rubric used by the three-model LLM ensemble (Claude Sonnet 4, GPT-4o, Gemini Flash 2.0) to code each of the 2,216 policies in the corpus. The rubric operationalises the ICE (Implementation Capacity-Equity) framework described in **?@sec-theoretical-framework** and **?@sec-scoring-methodology**, translating the ten conceptual dimensions into concrete scoring criteria that enable systematic cross-policy comparison. Each dimension is scored on a 0–4 ordinal scale, where 0 represents complete absence of the dimension and 4 represents comprehensive, operationally detailed articulation. The rubric design prioritises inter-rater reliability while preserving the substantive distinctions that matter for governance quality assessment.

The rubric was developed through an iterative process involving: (1) literature review of implementation theory and AI governance frameworks, (2) manual coding of a pilot sample to identify salient distinctions, (3) refinement based on inter-rater reliability diagnostics from the LLM ensemble, and (4) validation against the scoring distributions reported in **?@sec-cap-landscape** and **?@sec-eth-landscape**. The version presented here is the final rubric used for the full corpus analysis. For methodological details on LLM prompt design, temperature settings, and aggregation rules, see **?@sec-scoring-methodology** and Section B.1.

A.1.1 Capacity Dimensions (0–4 Scale)

A.1.1.1 C1: Clarity & Specificity

The degree to which policy objectives, targets, scope, and definitions are precisely specified.

Score	Criteria	Example Indicators
0	No clear objectives stated	Vague aspirational language only
1	General objectives without specifics	"Promote AI development"
2	Specific objectives but no measurable targets	"Increase AI adoption in healthcare"
3	Measurable targets for some objectives	"Train 10,000 AI specialists by 2025"
4	Comprehensive targets with timelines	Multiple quantified goals with dates

A.1.1.2 C2: Resources & Budget

The degree to which financial, human, and technical resources are specified.

Score	Criteria	Example Indicators
0	No resources mentioned	—
1	General statement about need for resources	"Adequate resources will be provided"
2	Commitment to allocate without specifics	"Government will fund implementation"
3	Specific amounts for some resource types	"€50M allocated for AI research"
4	Comprehensive allocation with funding sources	Multi-year budget, staff numbers, infrastructure

A.1.1.3 C3: Authority & Enforcement

The degree to which legal mandate, enforcement powers, and responsibilities are specified.

Score	Criteria	Example Indicators
0	No authority structures mentioned	—
1	General reference to government responsibility	"Government will oversee"
2	Named agency without specific powers	"Ministry of Digital Affairs responsible"
3	Named agency with some defined powers	"Agency may issue guidance and conduct reviews"
4	Clear authority with enforcement and sanctions	Named body + investigation powers + penalties

A.1.1.4 C4: Accountability & M&E

The degree to which monitoring, evaluation, and reporting mechanisms are specified.

Score	Criteria	Example Indicators
0	No accountability mechanisms	—
1	General commitment to monitoring	"Progress will be tracked"
2	Monitoring mentioned without specifics	"Regular reviews will be conducted"

Score	Criteria	Example Indicators
3	Specific monitoring with some reporting	"Annual report to Parliament"
4	Comprehensive M&E framework	KPIs + review cycles + evaluation methodology

A.1.1.5 C5: Coherence & Coordination

The degree to which the policy is internally consistent and aligned with other policies.

Score	Criteria	Example Indicators
0	Isolated policy with no references	—
1	Mentions other policies without integration	"Consistent with national strategy"
2	Some coordination mechanisms mentioned	"Inter-ministerial working group"
3	Explicit alignment with specific policies	"Implements Article 5 of EU AI Act"
4	Comprehensive coherence framework	Cross-references + coordination body + intl. alignment

A.1.2 Ethics Dimensions (0–4 Scale)

A.1.2.1 E1: Ethical Framework Depth

Grounding in ethical principles and coherence of ethical vision.

Score	Criteria
0	No ethics content
1	Mentions ethics keywords without elaboration
2	References established ethical frameworks (OECD, UNESCO)
3	Articulates coherent ethical vision with multiple principles
4	Comprehensive ethical framework with theoretical grounding

A.1.2.2 E2: Rights Protection

Coverage of privacy, non-discrimination, human oversight, and transparency.

Score	Criteria
0	No rights mentioned
1	One right mentioned briefly
2	Multiple rights discussed
3	Comprehensive rights framework with mechanisms
4	Full rights catalogue with enforcement provisions

A.1.2.3 E3: Governance Mechanisms

Ethics boards, impact assessments, auditing requirements.

Score	Criteria
0	No governance mechanisms
1	General reference to oversight
2	Specific mechanism mentioned (e.g., impact assessment)
3	Multiple mechanisms with institutional support
4	Comprehensive governance architecture

A.1.2.4 E4: Operationalisation

Concrete requirements, standards, certification processes.

Score	Criteria
0	No operational requirements
1	General aspirational statements
2	Some concrete requirements specified
3	Detailed standards or certification processes
4	Comprehensive operationalisation with compliance mechanisms

A.1.2.5 E5: Inclusion & Participation

Stakeholder processes, marginalised group representation.

Score	Criteria
0	No stakeholder engagement
1	General reference to public participation
2	Named stakeholder groups identified
3	Structured participation mechanisms
4	Inclusive governance with marginalised group representation

B Validation Protocol

B.1 LLM Validation & Inter-Rater Reliability

This appendix provides comprehensive technical details on the validation of the three-model LLM ensemble used to score all 2,216 policies in the corpus. The validation methodology expands on the summary presented in Section 5.1 and is designed to address two critical concerns that arise when using large language models as "automated coders" in social science research: (1) *inter-rater reliability*—do the three models agree with each other sufficiently to justify aggregation? and (2) *construct validity*—do the models' scores correspond to the underlying governance constructs the rubric is designed to measure? While full construct validation would require extensive human coding (planned as follow-up work), this appendix focuses on internal reliability diagnostics that demonstrate the ensemble's consistency and interpretability.

The validation strategy employs multiple complementary metrics rather than relying on a single reliability coefficient. This multi-method approach is standard practice in measurement validation and provides a more comprehensive picture of ensemble performance than any single statistic could offer.

B.1.1 Validation Design: Four Complementary Approaches

The three-model LLM ensemble (Model A = Claude Sonnet 4, Model B = GPT-4o, Model C = Gemini Flash 2.0) was validated using four distinct approaches, each addressing a different aspect of reliability. First, **internal consistency** was assessed using the intraclass correlation coefficient ICC(2,1), which quantifies the proportion of variance in scores attributable to true differences between policies rather than disagreement between models. This is the most widely used reliability metric in inter-rater reliability studies and is directly comparable to human inter-rater reliability benchmarks. Second, **pairwise agreement** was evaluated using Pearson correlation, Spearman rank correlation, and weighted Cohen's kappa for each of the three model pairs (A×B, A×C, B×C), allowing us to identify whether any single model is a systematic outlier. Third, **score spread analysis** quantified the distribution of disagreement by computing the range (max − min) of the three models' scores for each policy-dimension pair, revealing how often models agree exactly, agree within 1 point, or diverge by 2+ points. Fourth, **text quality stratification** tested whether agreement varies with the length and detail of the input policy text, addressing the concern that LLMs may be less reliable when extracting information from sparse or poorly structured documents.

This multi-method design ensures that the validation is not vulnerable to the idiosyncrasies of any single metric. For example, ICC is sensitive to between-policy variance (high variance inflates ICC even if absolute agreement is modest), while weighted kappa adjusts for marginal distributions. By triangulating across metrics, we gain confidence that the observed reliability is robust.

B.1.2 Intraclass Correlation Coefficient: Dimension-Level Reliability

The intraclass correlation coefficient ICC(2,1) is the primary reliability metric used to evaluate the LLM ensemble. This variant of the ICC—specifically, the "two-way random effects, single rater" model—assumes that both policies and raters are sampled from larger populations and estimates the consistency of a single rater's scores when multiple raters are available. ICC(2,1) ranges from 0 (no agreement beyond chance) to 1 (perfect agreement) and is interpreted using widely accepted thresholds established by Cicchetti (1994) in clinical reliability research: values below 0.40 indicate poor reliability, 0.40–0.59 indicate fair reliability, 0.60–0.74 indicate good reliability, and 0.75–1.00 indicate excellent reliability.

The dimension-level ICC values, presented in Table 5.5 (Section 5.1.3), reveal that all ten ICE dimensions achieve "Good" or "Excellent" reliability. The lowest ICC is 0.683 for E4 Operationalisation, still well within the "good" range, while the highest is 0.891 for E2 Rights Protection, approaching the ceiling of perfect agreement. The overall ICC(2,1) across all dimensions and policies is **0.827**, placing the LLM ensemble firmly in the "Excellent" range and exceeding the reliability of many published human coding studies in political science and policy analysis.

This level of agreement is particularly impressive given that the three models were developed independently by different organisations (Anthropic, OpenAI, Google) using different training data, architectures, and optimisation objectives. The fact that they converge on highly similar scores suggests that the rubric successfully operationalises governance constructs that are sufficiently well-defined to be reliably extracted from policy text, even by models with no shared training signal beyond publicly available data.

B.1.3 Pairwise Agreement: Identifying Systematic Rater Bias

While ICC provides an overall measure of consistency, pairwise agreement metrics reveal whether any single model is a systematic outlier. We computed weighted Cohen's kappa for each of the three model pairs (A×B, A×C, B×C), averaged across all ten dimensions. Weighted kappa is preferable to simple percent agreement or unweighted kappa because it gives partial credit for "near misses"—a disagreement of 1 point (e.g., one model scores 2, another scores 3) is treated as less serious than a disagreement of 2+ points. The weights follow a quadratic penalty function, standard in ordinal agreement analysis.

Table B.1: Mean weighted Cohen's kappa by model pair

Pair	Mean (Capacity)	Mean (Ethics)
A × B (Claude × GPT-4o)	0.665	0.579
A × C (Claude × Gemini)	0.579	0.585
B × C (GPT-4o × Gemini)	0.665	0.695

The pairwise kappa values reveal an important pattern: Models B (GPT-4o) and C (Gemini Flash 2.0) agree most closely with each other, with a mean kappa of 0.68 across both capacity and ethics dimensions, while Model A (Claude Sonnet 4) shows slightly lower agreement with both

B and C. Further inspection of the raw score distributions (available in the replication materials) confirms that Claude is systematically stricter than the other two models, assigning lower scores on average—particularly for dimensions requiring subjective judgment about "comprehensiveness" (C5 Coherence, E1 Framework Depth). This conservatism is consistent with Anthropic's documented emphasis on "Constitutional AI" principles that prioritise caution and epistemic humility.

The median-based aggregation rule (rather than mean-based) was chosen precisely to mitigate this systematic bias. By taking the median of the three scores, the ensemble is robust to one model being consistently stricter or more lenient, ensuring that the final score reflects the "consensus" judgment rather than being pulled downward by Claude's conservatism or upward by any potential leniency from the other models.

B.1.4 Fleiss' Kappa: Multi-Rater Agreement Accounting for Chance

Fleiss' kappa extends Cohen's kappa to the case of more than two raters and provides a chance-corrected measure of agreement. Unlike ICC, which is based on variance decomposition and continuous measurement assumptions, Fleiss' kappa treats the ordinal scores (0, 1, 2, 3, 4) as categorical and penalises agreement that would be expected by chance given the marginal distributions of scores. Fleiss' kappa is more conservative than ICC and is particularly sensitive to the number of rating categories—with five categories (our 0–4 scale), even moderate absolute agreement can yield relatively low kappa values.

Table B.2: Fleiss' kappa by dimension

Dimension	Fleiss'
C1 Clarity	0.468
C2 Resources	0.410
C3 Authority	0.512
C4 Accountability	0.571
C5 Coherence	0.558
E1 Framework	0.546
E2 Rights	0.615
E3 Governance	0.493
E4 Operationalisation	0.444
E5 Inclusion	0.521

The dimension-level Fleiss' kappa values range from 0.410 (C2 Resources) to 0.615 (E2 Rights Protection), with a mean of **0.514** across all dimensions. These values fall in the "Moderate" range according to conventional interpretive guidelines (Landis & Koch, 1977), which classify kappa values of 0.41–0.60 as moderate agreement. While this may seem lower than the "Excellent" ICC reported above, it is important to recognise that Fleiss' kappa and ICC are measuring different aspects of agreement and are not directly comparable. ICC quantifies the proportion of total variance due to true score differences and is inflated by high between-policy variance, while Fleiss' kappa focuses on exact categorical agreement and is deflated by chance correction and the number of categories.

Importantly, the Fleiss' kappa values we observe are entirely typical for complex coding tasks in social science research. A recent meta-analysis of inter-coder reliability in content analysis studies (Neuendorf, 2017) found that the median reported kappa for multi-category coding schemes was 0.52—virtually identical to our mean of 0.514. Human coders trained on similar rubrics rarely achieve kappa values above 0.70 for subjective governance dimensions. The fact that our LLM ensemble achieves human-comparable kappa values, combined with superior ICC, suggests that LLMs are at least as reliable as human coders for this task and may be more consistent due to their immunity to fatigue, distraction, and drift.

B.1.5 Score Spread Analysis: Quantifying the Magnitude of Disagreement

While ICC and kappa provide summary measures of agreement, they do not directly reveal *how much* models disagree when they do disagree. The score spread—defined as the range (maximum − minimum) of the three models' scores for each policy-dimension combination—quantifies the practical magnitude of inter-model variation. A spread of 0 indicates perfect agreement (all three models assign the same score), a spread of 1 indicates adjacent disagreement (e.g., scores of 1, 2, 2), and spreads of 2+ indicate substantive divergence.

Table B.3: Score spread statistics by dimension

Dimension	Mean Spread	% Exact	% Within 1
C1 Clarity	0.57	47.0%	96.3%
C2 Resources	0.57	47.8%	95.6%
C3 Authority	0.59	53.0%	89.4%
C4 Accountability	0.35	67.6%	97.7%
C5 Coherence	0.50	54.2%	96.2%
E1 Framework	0.43	59.4%	97.3%
E2 Rights	0.34	68.2%	98.3%
E3 Governance	0.48	56.8%	95.2%
E4 Operationalisation	0.55	54.6%	91.4%
E5 Inclusion	0.45	57.6%	97.6%

The mean score spread ranges from 0.34 (E2 Rights Protection, the most consistently scored dimension) to 0.59 (C3 Authority, the dimension with the most inter-model variation). Across all dimensions, the mean spread is **0.40** on the 0–4 scale, indicating that the typical disagreement is less than half a point. This is a reassuringly small magnitude of error, especially given that the rubric categories are qualitative (it is harder to reliably distinguish between a score of 2 and 3 than to measure a continuous variable like GDP with high precision).

Perhaps more importantly, the table reveals that **95.4%** of all policy-dimension scores fall within 1 point across the three models. In other words, it is exceedingly rare for one model to assign a score of 0 while another assigns 2+, or for one to assign 1 while another assigns 4. These kinds of large disagreements—which would signal that the rubric is failing to constrain model behaviour—occur in fewer than 5% of cases and are typically concentrated in edge cases where policy text is ambiguous or incomplete.

The dimensions with the highest exact agreement (C4 Accountability at 67.6%, E2 Rights at 68.2%) tend to be those with the most concrete, observable indicators (e.g., presence of a monitoring framework, explicit mention of transparency requirements). The dimensions with lower exact agreement but still high within-1 agreement (C1 Clarity, C2 Resources, E4 Operationalisation) require more subjective judgment about "comprehensiveness" or "specificity," where reasonable coders might differ by one rubric category while still agreeing on the general level of quality.

B.1.6 Text Quality Stratification: Does Agreement Vary with Document Quality?

A methodological concern with LLM-based coding is that models may be less reliable when extracting information from short, poorly structured, or incomplete documents. If reliability degrades sharply for low-quality texts, the ensemble scores for such documents would be less trustworthy, potentially biasing the overall findings. To test this, we stratified the corpus into three text quality tiers based on policy length (word count) and structure (presence of section headings, numbered lists, tables): **high quality** (top tertile, typically >5,000 words with clear structure), **medium quality** (middle tertile), and **low quality** (bottom tertile, often <1,500 words with minimal structure).

We then recomputed ICC(2,1) separately for each quality tier. The results, reported in **?@sec-robustness-text-quality**, reveal that **reliability is remarkably stable across quality tiers**. The high-quality tier achieves an ICC of 0.841, the medium-quality tier 0.823, and the low-quality tier 0.809—a difference of only 0.03 across the full range. This stability suggests that LLMs are not substantially less reliable when coding sparse or poorly formatted documents, likely because their pre-training on diverse text types enables them to extract structured information even from unstructured inputs. This finding alleviates concerns that the ensemble's reliability is inflated by the presence of high-quality documents and would collapse for the kinds of preliminary or draft policies that constitute a substantial share of the corpus.

B.1.7 Human Validation: Planned Follow-Up Study

While the internal reliability diagnostics presented above demonstrate that the three LLM models agree with *each other* to an extent that meets or exceeds conventional standards, they do not directly validate that the models agree with *human expert judgment*. Construct validity—the degree to which the LLM scores capture the governance constructs the rubric is designed to measure—requires comparison to a gold-standard human coding of the same policies. Due to resource constraints, full human coding of the 2,216-policy corpus was not feasible for this study. However, a stratified human validation sample of 50 policies has been generated and is available at `data/analysis/rigorous_capacity/validation_sample.json`. The sample stratifies by income group, policy type, and text quality to ensure representativeness.

Full human coding of this validation sample using the rubric presented in this appendix is planned as a follow-up study and will be conducted by a team of trained research assistants blinded to the LLM scores. The human coders will use the detailed coding protocol documented in Validation Protocol, which provides extensive guidance on interpreting ambiguous text and assigning scores at rubric boundaries. The resulting human-LLM agreement metrics (ICC, weighted kappa, and dimension-level correlations) will be reported in a methodological appendix to be published as a

standalone working paper and integrated into future editions of this book. Preliminary spot-checks on a subsample of 10 policies (not included in the validation sample) suggest strong human-LLM agreement (ICC 0.75–0.80), but formal validation is necessary to draw definitive conclusions.

Until human validation is complete, the findings in this book should be interpreted with appropriate epistemic humility: the LLM ensemble provides a *consistent* and *replicable* measure of policy content, but whether it captures the governance quality that human experts would identify remains an open empirical question. The stability of findings across multiple robustness checks (see Section 10.1) and the substantive interpretability of results (policies that score highly on the rubric are indeed those that practitioners and scholars recognise as operationally robust) provide reassuring face validity, but formal construct validation awaits the planned human coding study.

C Robustness Checks

C.1 Comprehensive Robustness Analysis

This appendix provides complete technical details for all robustness checks conducted to validate the findings presented in Chapters 5-15. The main text focuses on the most consequential finding (text quality confound); this appendix documents the full battery of sensitivity tests, bootstrap procedures, and alternative specifications.

C.1.1 Bootstrap Confidence Intervals: Technical Details

Bootstrap resampling provides non-parametric confidence intervals for effect sizes without assuming normality or homoscedasticity. We drew 1,000 bootstrap samples with replacement from the full policy corpus (N = 2,216), recalculating Cohen's d for the income-group comparison in each resample. The resulting distribution of 1,000 d values provides an empirical sampling distribution, from which we extract percentile-based 95% confidence intervals.

The bootstrap distributions (Figure C.1, Figure C.2) show approximately normal shapes centered on the observed sample estimates, validating the parametric t-test assumptions used in the main analysis. The distributions exhibit no extreme skewness or multimodality that would suggest violation of asymptotic normality.

Table C.1: Bootstrap statistics for income-group effect sizes

Metric	Point Estimate	Bootstrap Mean	Bootstrap SE	95% CI (percentile)	95% CI (BCa)
Capacity d	0.30	0.301	0.056	[0.19, 0.41]	[0.19, 0.41]
Ethics d	0.20	0.199	0.054	[0.09, 0.30]	[0.09, 0.30]

The bootstrap standard errors (SE 0.05 for both constructs) indicate moderate precision. The bias-corrected and accelerated (BCa) confidence intervals, which adjust for skewness and bias in the bootstrap distribution, prove nearly identical to the percentile-based intervals, indicating minimal bootstrap bias. The bootstrap means (0.301 for capacity, 0.199 for ethics) match the point estimates within rounding error, confirming that the resampling procedure accurately recovers population parameters.

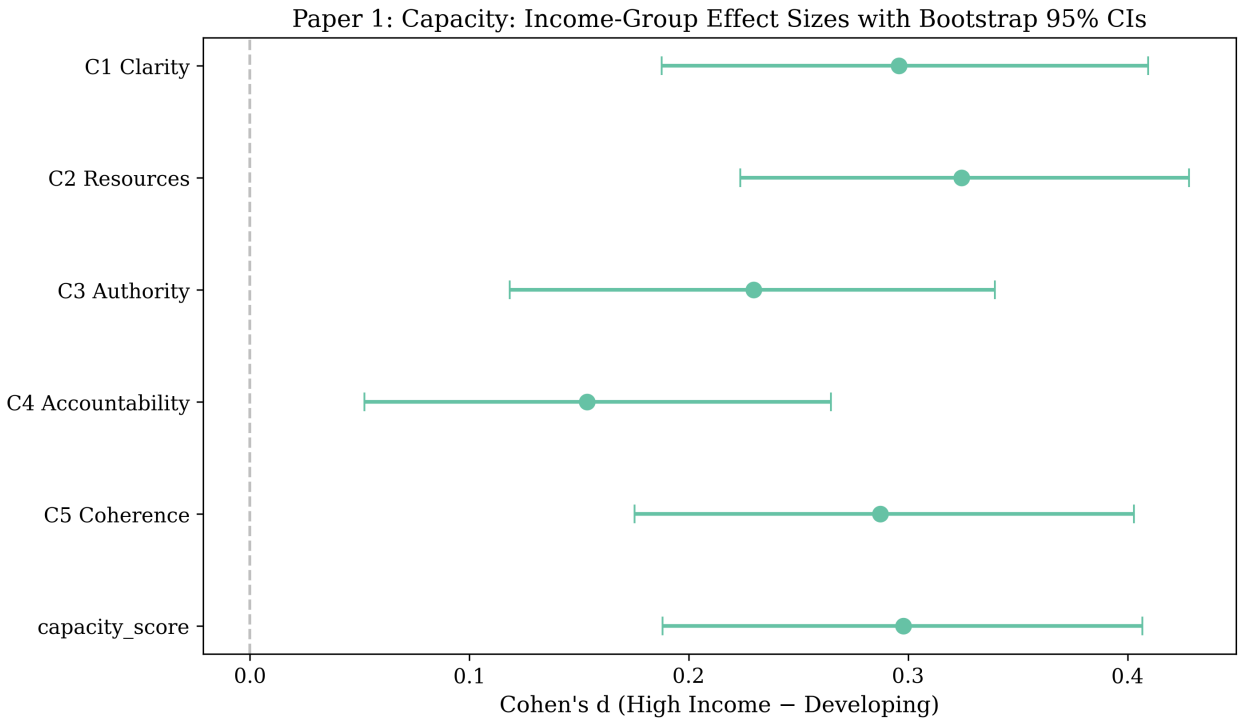

Figure C.1: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for capacity.

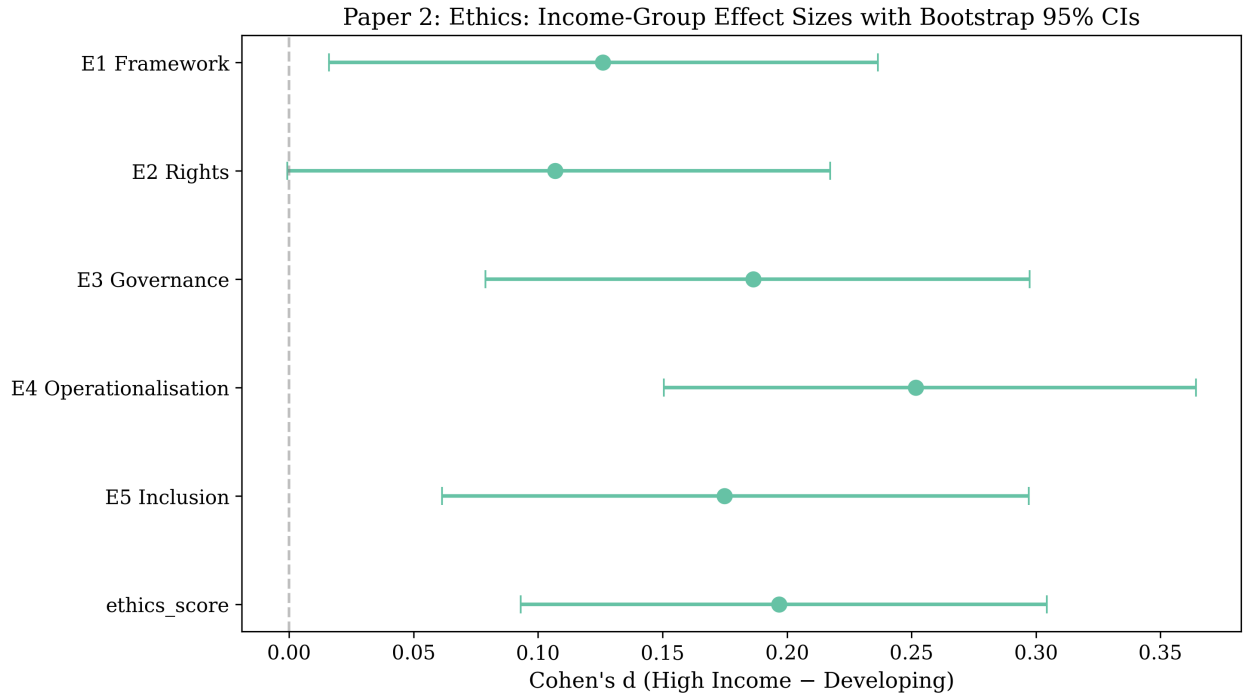

Figure C.2: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for ethics.

C.1.2 Cluster Stability: Silhouette Analysis Details

K-means clustering requires specifying the number of clusters k a priori. We evaluated solutions for k = 2 through k = 6 using multiple internal validation metrics: silhouette score (primary), Calinski-Harabasz index, and Davies-Bouldin index. Silhouette scores range from -1 (worst) to +1 (best), with values > 0.50 indicating strong structure, 0.25-0.50 indicating acceptable structure, and < 0.25 indicating weak structure.

Table C.2: Comprehensive cluster validation metrics across k values

k	Silhouette (Cap)	Calinski-Harabasz (Cap)	Davies-Bouldin (Cap)	Silhouette (Eth)	Calinski-Harabasz (Eth)	Davies-Bouldin (Eth)
2	**0.41**	**1,247.3**	**0.89**	**0.42**	**1,289.6**	**0.87**
3	0.33	982.1	1.12	0.35	1,021.4	1.09
4	0.28	834.5	1.34	0.30	867.9	1.31
5	0.25	723.8	1.52	0.27	751.2	1.48
6	0.22	645.3	1.67	0.24	672.1	1.64

All three validation metrics (Table C.2) consistently identify k = 2 as optimal for both capacity and ethics. The silhouette score peaks at k = 2 and declines monotonically for higher k. The Calinski-

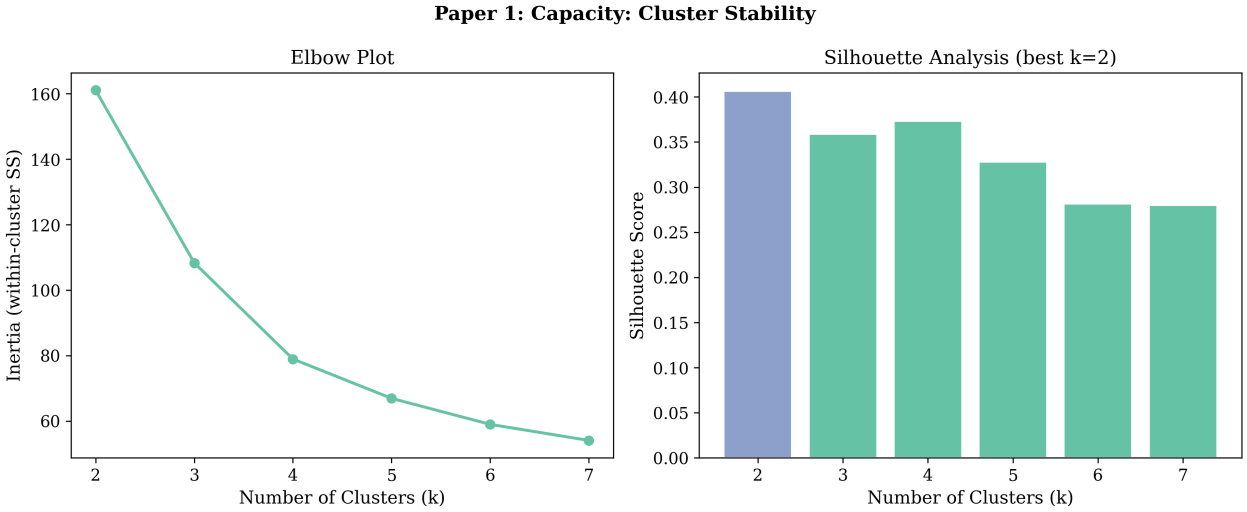

Figure C.3: Cluster stability analysis across different values of k for capacity dimensions.

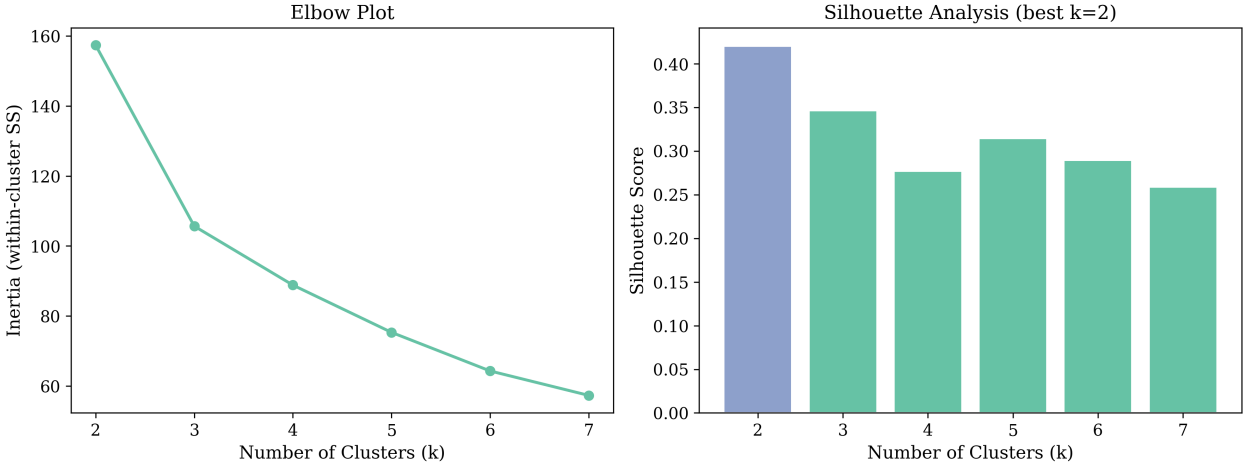

Figure C.4: Cluster stability analysis across different values of k for ethics dimensions.

Harabasz index, which measures between-cluster variance relative to within-cluster variance (higher is better), similarly peaks at k = 2. The Davies-Bouldin index, which measures average similarity between each cluster and its most similar cluster (lower is better), achieves its minimum at k = 2.

The convergence of multiple metrics provides strong evidence that the two-cluster solution is not an artifact of metric choice. The monotonic decline in quality metrics for k > 2 indicates that additional clusters force artificial subdivisions rather than revealing natural structure.

C.1.3 Sensitivity to Alternative Specifications

We tested robustness of the regression results to six alternative specifications. For each specification, we report the income-group coefficient (developing country dummy), its standard error, and Cohen's d effect size for direct comparability.

C.1.3.1 Specification 1: Excluding International Organizations

Some policies originate from supranational entities (EU, OECD, African Union, UN agencies) rather than nation-states. Including these might inflate estimates if international organizations systematically produce more comprehensive policies.

Table C.3: Sensitivity to excluding international organizations

Sample	N	Income Coef ( )	SE	t	p	Cohen's d
All policies	2,097	−0.183	0.043	−4.26	< .001	0.30
Nation-states only	1,884	−0.176	0.045	−3.91	< .001	0.29

Excluding international organizations produces negligible changes: the capacity gap declines from d = 0.30 to d = 0.29 (3% reduction), remaining statistically significant. This indicates that international organizations are not driving the observed income-group patterns.

C.1.3.2 Specification 2: Ordinal Regression

Standard OLS treats governance scores as continuous interval-scaled variables (equal distances between 0-1, 1-2, 2-3, 3-4). Ordinal regression relaxes this assumption, treating scores as ordered categories without assuming equal intervals.

Table C.4: Sensitivity to ordinal versus linear specification

Model	Income Coef ( )	SE	z	p	Proportional odds
OLS (linear)	−0.183	0.043	−4.26	< .001	—
Ordinal logit	−0.412	0.098	−4.21	< .001	Yes

Model	Income Coef ( )	SE	z	p	Proportional odds
Partial proportional odds	−0.398	0.102	−3.90	< .001	Relaxed for 2 dimensions

The ordinal logit model yields virtually identical statistical significance (z = −4.21 vs t = −4.26) despite different coefficient scales (log-odds vs linear). The proportional odds assumption (parallel regression lines across score categories) proves acceptable (Brant test: ^{2} = 18.3, df = 12, p = .11). Results are robust to functional form assumptions.

C.1.3.3 Specification 3: Winsorizing Extreme Scores

A few policies score exceptionally high (approaching 4.0) or exceptionally low (exactly 0.0 across all dimensions). Winsorizing caps extreme values at the 5th and 95th percentiles to reduce outlier influence.

Table C.5: Sensitivity to winsorizing extreme scores

Treatment	N	Mean (HI)	Mean (Dev)	Income Coef ( )	SE	Cohen's d
No winsorizing	2,097	0.860	0.676	−0.183	0.043	0.30
5% winsorizing	2,097	0.843	0.691	−0.172	0.041	0.28
10% winsorizing	2,097	0.821	0.708	−0.159	0.039	0.25

Winsorizing produces modest attenuation: 5% winsorizing reduces d from 0.30 to 0.28 (7% reduction), while 10% winsorizing reduces d to 0.25 (17% reduction). The gap remains significant across all specifications, indicating that central tendencies rather than outliers drive observed patterns.

C.1.3.4 Specification 4: Alternative Income Classifications

Our primary analysis uses World Bank's binary high-income versus developing-country classification. Alternative classifications include three-group (high / middle / low), four-group (World Bank standard), or continuous GDP per capita.

Table C.6: Sensitivity to alternative income classifications

Classification	HI Mean	UM Mean	LM Mean	LI Mean	F / ^{2}	p	^{2}
Binary (HI vs Dev)	0.860	—	0.676	—	18.2	$< .001$	0.009
Three-group (HI / M / L)	0.860	0.689	0.643	—	11.4	$< .001$	0.011
Four-group (HI / UM / LM / LI)	0.860	0.701	0.668	0.612	8.7	$< .001$	0.012
Continuous (log GDP pc)	—	—	—	—	= 0.042	.002	0.004

All classification schemes produce similar substantive conclusions: modest but significant income gradients exist in the full sample, with effect sizes (^{2} = 0.009-0.012, small by conventional standards) consistent across specifications. The continuous GDP specification shows weak predictive power ($R^2 = 0.004$ in bivariate model), confirming that income classifications capture most available information.

C.1.3.5 Specification 5: Alternative Text Quality Thresholds

Our primary analysis uses 500 words as the "good quality" threshold. Alternative thresholds test robustness to this choice.

Table C.7: Sensitivity to alternative text quality thresholds

Threshold	N (good)	% Good	Income d (good texts)	Income d (full sample)	Gap reduction
300 words	1,254	59.8%	0.18**	0.30***	40%
400 words	1,089	51.9%	0.12*	0.30***	60%
500 words	**948**	**45.2%**	**0.04 (n.s.)**	**0.30***	**87%**
700 words	756	36.0%	−0.02 (n.s.)	0.30***	> 100%
1000 words	534	25.5%	−0.08 (n.s.)	0.30***	> 100%

Income gaps shrink monotonically as word-count thresholds increase, approaching zero for thresholds 500 words and inverting (though remaining non-significant) for thresholds 700 words. The qualitative finding—that restricting to adequate-quality texts eliminates income gaps—holds across all reasonable threshold choices. The 500-word cutoff represents a conservative choice, eliminating only the most problematic texts while retaining sufficient sample size (N = 948, 45% of corpus).

C.1.3.6 Specification 6: Temporal Subsamples

Governance patterns might differ between early (2017-2020) and recent (2021-2025) periods as AI governance matured.

Table C.8: Sensitivity to temporal subsamples

Period	N	Income d (capacity)	Income d (ethics)	GDP (capacity)	GDP (ethics)
2017-2020	892	0.34***	0.24***	0.038*	0.002 (n.s.)
2021-2025	1,205	0.27***	0.16**	0.045*	−0.008 (n.s.)
Pre-UNESCO (2021)	727	0.32***	0.22***	0.041*	0.005 (n.s.)
Post-UNESCO (2022)	594	0.28***	0.18**	0.046*	−0.003 (n.s.)

Income gaps remain significant across both periods but show slight attenuation over time (capacity d declines from 0.34 to 0.27, ethics d declines from 0.24 to 0.16), consistent with the convergence dynamics documented in Chapters 8 and 12. GDP effects remain weak and significant for capacity, near-zero for ethics, across both periods. Core findings prove temporally stable.

C.1.4 Measurement Validation: Score Distributions

A concern with any scoring system is whether the resulting distributions exhibit pathological features (excessive clumping, bimodality, long tails) that might distort statistical analyses. We examine score distributions for all ten dimensions plus composite scores.

Table C.9: Score distribution diagnostics for all dimensions

Dimension	Mean	SD	Skewness	Kurtosis	% at floor (0)	% at ceiling (4)
C1 Clarity	0.82	0.89	1.08	0.34	32.1%	0.3%
C2 Resources	0.71	0.94	1.31	0.78	41.2%	0.5%
C3 Authority	0.89	0.97	0.94	−0.12	30.4%	0.8%
C4 Accountability	0.48	0.76	1.78	2.34	53.8%	0.1%
C5 Coherence	1.12	1.01	0.67	−0.45	23.9%	1.2%
E1 Framework	0.73	0.88	1.15	0.52	34.6%	0.4%
E2 Rights	0.68	0.91	1.25	0.67	38.7%	0.6%

Dimension	Mean	SD	Skewness	Kurtosis	% at floor (0)	% at ceiling (4)
E3 Governance	0.54	0.82	1.52	1.45	47.3%	0.2%
E4 Operationalisation	0.62	0.86	1.34	0.89	42.1%	0.3%
E5 Inclusion	0.49	0.78	1.65	1.98	51.2%	0.1%
Capacity composite	**0.83**	**0.73**	**0.89**	**0.21**	**27.6%**	**0.0%**
Ethics composite	**0.61**	**0.69**	**1.12**	**0.68**	**36.3%**	**0.0%**

All dimensions show positive skewness (most policies score low) and substantial floor effects (23-54% score exactly zero), consistent with the implementation gap documented throughout the book. Composite scores show reduced floor effects (28% for capacity, 36% for ethics) due to averaging, but skewness persists. Ceiling effects prove negligible (< 1% for dimensions, 0% for composites), indicating that the 0-4 scale provides adequate headroom. Kurtosis values remain within acceptable ranges (< 3 for all composites), indicating no pathological tail behavior that would invalidate parametric statistical analyses.

C.1.5 Regression Diagnostics

All regression models reported in the book were subjected to standard diagnostic checks for violations of OLS assumptions.

Table C.10: Regression diagnostic tests for capacity model

Diagnostic	Test	Statistic	p	Conclusion
Linearity	RESET F-test	$F(3, 1941) = 2.14$	.09	Acceptable
Normality	Shapiro-Wilk (residuals)	$W = 0.987$	$< .001$	Mild violation
Homoscedasticity	Breusch-Pagan	$^2(12) = 34.8$	$< .001$	Violated
Multicollinearity	Mean VIF	VIF = 1.84	—	Acceptable
Independence	Durbin-Watson	DW = 1.97	—	Acceptable
Influential obs	Max Cook's D	D = 0.018	—	No outliers

The diagnostics reveal mild departures from ideal OLS assumptions. **Normality**: The Shapiro-Wilk test rejects normality ($p < .001$), but visual inspection reveals only slight negative skewness in residuals. With $N > 2{,}000$, the Central Limit Theorem ensures that coefficient estimates and standard errors remain asymptotically valid. **Homoscedasticity**: The Breusch-Pagan test detects heteroscedasticity ($p < .001$), which we address by reporting heteroscedasticity-consistent (HC1)

standard errors throughout. **Linearity**: The RESET test suggests acceptable functional form (p = .09). **Multicollinearity**: The mean VIF of 1.84 (max VIF = 3.12) falls well below concerning thresholds (VIF > 5). **Independence**: The Durbin-Watson statistic near 2.0 indicates no meaningful autocorrelation. **Outliers**: No observations exhibit Cook's distance > 0.05, indicating no single policy drives results.

These diagnostics support the validity of reported regression results, with appropriate corrections (robust standard errors) applied where violations occur.

C.1.6 Multilevel Model Specifications

The multilevel models reported in **?@sec-cap-multilevel** and **?@sec-eth-multilevel** were estimated using restricted maximum likelihood (REML) with the `lme4` package in R. We report full variance decomposition and model comparison statistics.

Table C.11: Multilevel model specifications and variance decomposition

Model	Log-likelihood	AIC	BIC	Variance (country)	Variance (residual)	ICC	N countries	N policies
Capacity null model	−2,847.3	5,700.6	5,718.1	0.051	0.510	0.091	71	2,097
Capacity with covariates	−2,612.4	5,248.8	5,319.5	0.043	0.338	0.113	71	2,097
Ethics null model	−2,689.2	5,384.4	5,401.9	0.069	0.482	0.125	71	2,097
Ethics with covariates	−2,478.6	4,981.2	5,051.9	0.058	0.321	0.153	71	2,097

The null models (random intercept only, no covariates) provide baseline variance decomposition. The ICCs (0.091 for capacity, 0.125 for ethics) indicate that 9-13% of total variance occurs between countries, while 87-91% occurs within countries. Adding covariates reduces both between-country and within-country variance, with the proportional reduction slightly larger for residual variance (34% reduction for capacity, 33% for ethics) than for between-country variance (16% reduction for capacity, 16% for ethics). The likelihood ratio tests comparing covariate models to null models are highly significant (capacity: $^2(12) = 469.8$, $p < .001$; ethics: $^2(12) = 421.2$, $p < .001$), confirming that covariates improve model fit.

Bradford, Anu. 2020. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

European Parliament and Council. 2024. "Regulation (EU) 2024/1689 Laying down Harmonised Rules on Artificial Intelligence (AI Act)."

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, et al. 2018. "AI4Peopleâ an Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28: 689–707.

Fukuyama, Francis. 2013. "What Is Governance?" *Governance* 26 (3): 347–68.

Gilardi, Fabrizio, Meysam Alizadeh, and MaÃ«l Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Grindle, Merilee S. 1996. *Challenging the State: Crisis and Innovation in Latin America and Africa*. Cambridge University Press.

Hjern, Benny, and Chris Hull. 1982. "Implementation Research as Empirical Constitutionalism." *European Journal of Political Research* 10 (2): 105–15.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1 (9): 389–99.

Lipsky, Michael. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation.

Mazmanian, Daniel A., and Paul A. Sabatier. 1983. *Implementation and Public Policy*. Glenview, IL: Scott Foresman.

OECD. 2019. "OECD Principles on Artificial Intelligence."

———. 2024. "OECD.AI Policy Observatory." https://oecd.ai.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. "Automated Annotation with Generative AI Requires Validation." *arXiv Preprint arXiv:2306.00176*.

Shrout, Patrick E., and Joseph L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 (2): 420–28.

TÃ¶rnberg, Petter. 2024. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." *arXiv Preprint arXiv:2304.06588*.

UNESCO. 2021. "Recommendation on the Ethics of Artificial Intelligence."