Global Observatory of AI Ethics Governance

Measuring Normative Commitments Across 2,100+ Policies

Lucas Sempé

February 11, 2026

Table of contents

1 Global Observatory of AI Ethics Governance

Measuring Normative Commitments Across 2,100+ Policies

Preface

This book presents the first systematic global assessment of AI **ethics governance** — measuring the depth and specificity of normative commitments in AI policies worldwide.

Drawing on 2,100+ policies across 70+ jurisdictions, we score each policy on five ethics dimensions: framework depth, rights protection, participatory governance, operationalization, and inclusion. The analysis reveals a landscape where ethical commitments vary more within income groups than between them.

1.1 Key Findings

- **Ethics convergence**: Gaps narrowing over time as developing countries strengthen ethical frameworks
- **Horizontal diffusion**: Regional policy learning dominates over North-South transfer
- **Within-group inequality**: 99% of variation occurs within income groups
- **Text quality confound**: Apparent gaps disappear for well-documented policies

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), enabling policy analysis at unprecedented scale.

Citation: Sempé, L. (2026). *Global Observatory of AI Ethics Governance*. International Initiative for Impact Evaluation (3ie).

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

2 Introduction

2.1 Measuring Normative Commitments in AI Governance

> **Chapter summary.** This chapter introduces the central research question: How deeply and specifically do AI policies embed ethical commitments? We examine gaps in ethics governance measurement and preview our analytical framework.

2.1.1 Beyond Principle Proliferation

The AI ethics landscape has produced remarkable normative convergence. Jobin et al. (2019) identified recurring principles (transparency, fairness, accountability) across 84 ethics guidelines. Fjeld et al. (2020) found similar patterns in 36 prominent frameworks.

Yet convergence on *principles* does not guarantee *governance depth*. Do policies merely list ethical values, or do they specify rights protections, establish participatory mechanisms, operationalize principles into actionable requirements, and ensure inclusive stakeholder engagement?

This book measures ethics governance depth across five dimensions:

1. **E1 Framework Depth**: Specificity of ethical principles and value articulation
2. **E2 Rights Protection**: Safeguards for privacy, fairness, non-discrimination, due process
3. **E3 Participatory Governance**: Public consultation, multi-stakeholder processes, transparency
4. **E4 Operationalization**: Concrete requirements, compliance mechanisms, enforcement
5. **E5 Inclusion**: Representation of marginalized groups, accessibility, equity considerations

2.1.2 Key Findings Preview

Our analysis of 2,100+ policies reveals:

- **Low average depth**: Mean 1.39/4.0, indicating superficial ethical commitments in most policies
- **Convergence over time**: Ethics gaps narrowing faster than capacity gaps
- **Within-group inequality**: 99% of variation occurs within income groups
- **Text quality confound**: Apparent income gaps (d=0.20) disappear (d=-0.09) for well-documented policies
- **Horizontal diffusion**: Regional ethics norms spread through peer learning

2.1.3 Roadmap

- **Chapter 9** maps global ethics governance landscape
- **Chapter 10** examines determinants of ethics governance depth
- **Chapter 11** analyzes inequality patterns
- **Chapter 12** traces temporal dynamics and convergence
- **Chapter 14** presents robustness checks
- **Chapters 15-16** discuss implications and conclusions

The following chapters reveal ethics governance as a domain where developing countries increasingly match wealthy-country commitments.

3 Literature Review

3.1 Theoretical Foundations

> **Chapter summary.** We situate AI ethics governance measurement within three literatures: (1) the AI ethics principles mapping wave, (2) normative frameworks for AI governance, and (3) the gap between principles and practice documented across regulatory domains.

3.1.1 The AI Ethics Mapping Wave

Jobin, Ienca, and Vayena (2019) conducted the foundational ethics mapping, analyzing 84 AI guidelines and identifying convergence around transparency, justice/fairness, non-maleficence, responsibility, and privacy. Floridi et al. (2018) proposed the AI4People framework adding beneficence and autonomy. Fjeld et al. (2020) extended mapping to 36 prominent frameworks.

These studies established **what principles appear** but not **how deeply policies operationalize them**. As Hagendorff (2020) observed, "the ethics of AI ethics"—the gap between principles and practice—remains the field's pressing challenge.

3.1.1.1 From Principles to Governance

(**mittelstadt2019?**) distinguished **principle-based** from **practice-based** AI ethics. Principles articulate values (fairness, transparency); practice requires translating values into actionable requirements, compliance mechanisms, and enforcement procedures. Most policies remain principle-based.

(**rességuier2020?**) proposed measuring ethics "embeddedness"—the degree to which principles become operationalized through concrete requirements. Our framework measures precisely this embeddedness across five dimensions.

3.1.2 Normative Frameworks

3.1.2.1 Rights-Based Approaches

(**hildebrandt2019?**) grounds AI ethics in fundamental rights, arguing that algorithmic systems threaten privacy, equality, and due process. Our **E2 Rights Protection** dimension captures whether policies establish specific safeguards for these rights.

(**yeung2018?**) shows rights-based governance requires translating abstract rights into technical specifications—a challenging task most policies avoid. This explains the low E2 scores documented in **?@sec-ethics-landscape**.

3.1.2.2 Participatory Governance

(**rahwan2018?**) argues that AI governance requires democratic participation since algorithmic systems encode societal values. (**stilgoe2020?**) demonstrates that meaningful participation demands structured processes, not token consultation.

Our **E3 Participatory Governance** dimension measures whether policies establish multi-stakeholder processes, public consultation mechanisms, and transparency requirements enabling democratic oversight.

3.1.2.3 Operationalization Challenges

(**selbst2019?**) documents "fairness gerrymandering"—policies proclaiming commitment to fairness without specifying operational definitions or compliance methods. (**whittaker2018?**) shows this pattern extends to transparency, accountability, and ethics principles broadly.

Our **E4 Operationalization** dimension distinguishes policies that merely mention principles from those specifying concrete requirements, compliance procedures, and enforcement mechanisms.

3.1.3 The Governance Gap

Multiple studies document gaps between ethical commitments and implementation:

- (**mittelstadt2019?**): "AI ethics guidelines remain disconnected from practice"
- Hagendorff (2020): "Principles fail to translate into tangible governance"

- Jobin, Ienca, and Vayena (2019): "Convergence on principles masks divergence on implementation"

This gap motivates our measurement approach: scoring not what policies say but what they establish—the governance infrastructure translating principles into practice.

3.1.4 Ethics and Development

(**lee2018?**) argues developing countries face ethical "catch-up" challenges, adopting AI without governance frameworks. (**gwagwa2020?**) shows African countries increasingly develop indigenous ethics frameworks rather than importing Western principles.

(**muller2021?**) documents how UNESCO's AI Ethics Recommendation (2021) provides global framework respecting diverse values while establishing common standards. This creates opportunity for ethics convergence that our temporal analysis (Chapter 12) examines.

3.1.5 Measurement Challenges

Existing ethics assessments rely on binary presence/absence (Fjeld et al. 2020) or qualitative evaluation (**mittelstadt2019?**). Neither scales to comprehensive global measurement. Our LLM-based scoring enables assessment across 2,100+ policies, measuring ethics governance depth rather than mere principle mention.

3.1.6 Contribution

This book provides:

1. **Operationalized framework**: Five ethics dimensions distinguishing principle mention from governance depth
2. **Validated methodology**: LLM ensemble achieving ICC = 0.827
3. **Global assessment**: First comprehensive ethics governance measurement
4. **Convergence analysis**: Testing whether ethics gaps narrow over time

The following chapters reveal that ethics governance varies more within income groups than between them, and that developing countries increasingly match wealthy-country commitments.

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.

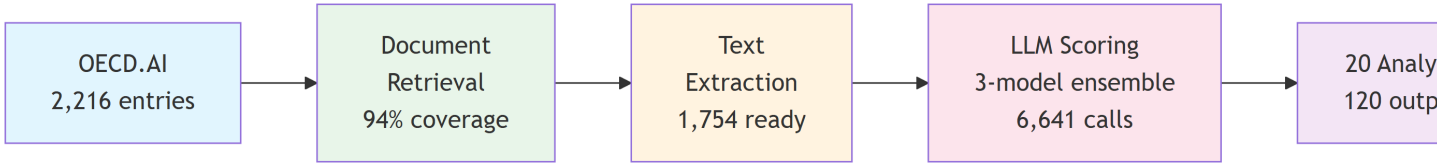

Figure 4.1: Analytical pipeline from corpus to results

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 Ethics Landscape

6.1 The Global Landscape of AI Ethics Governance

Chapter summary. This chapter mirrors the capacity landscape analysis for the five ethics dimensions. Ethics governance lags behind implementation capacity (mean 0.61 vs. 0.83), and 36.3% of all policies score exactly zero on ethics operationalisation.

6.1.1 Overall Score Distribution

The global landscape of AI ethics governance reveals a troubling pattern: policies worldwide prove substantially weaker on ethical principles than on implementation capacity. While capacity scores average 0.83 across the 2,216-policy corpus, ethics scores average just 0.61 — a gap of 0.22 points representing roughly one-fifth of the scale range. This disparity suggests that policymakers find it easier to specify implementation mechanisms (budgets, agencies, coordination procedures) than to articulate and operationalize ethical commitments. The floor effects visible in Figure 6.1 prove even more severe than for capacity, indicating that many policies contain minimal or zero ethical content despite addressing AI governance explicitly.

The distribution's right skew indicates that while a minority of policies demonstrate sophisticated ethics governance, the modal policy contains almost no ethical content. This creates a governance landscape where technical implementation details dominate and ethical principles remain aspirational rather than operationalized.

The ethics composite score averages **0.61/4.00** (SD = 0.62) — meaningfully lower than the capacity composite (0.83):

Table 6.1: Ethics dimension descriptive statistics

Dimension	Mean	SD	Median
E1 Ethical Framework Depth	0.67	0.75	0.00
E2 Rights Protection	0.55	0.66	0.00
E3 Governance Mechanisms	0.62	0.74	0.00
E4 Operationalisation	0.59	0.73	0.00
E5 Inclusion & Participation	0.65	0.73	0.00
Ethics composite	**0.61**	**0.62**	**0.40**

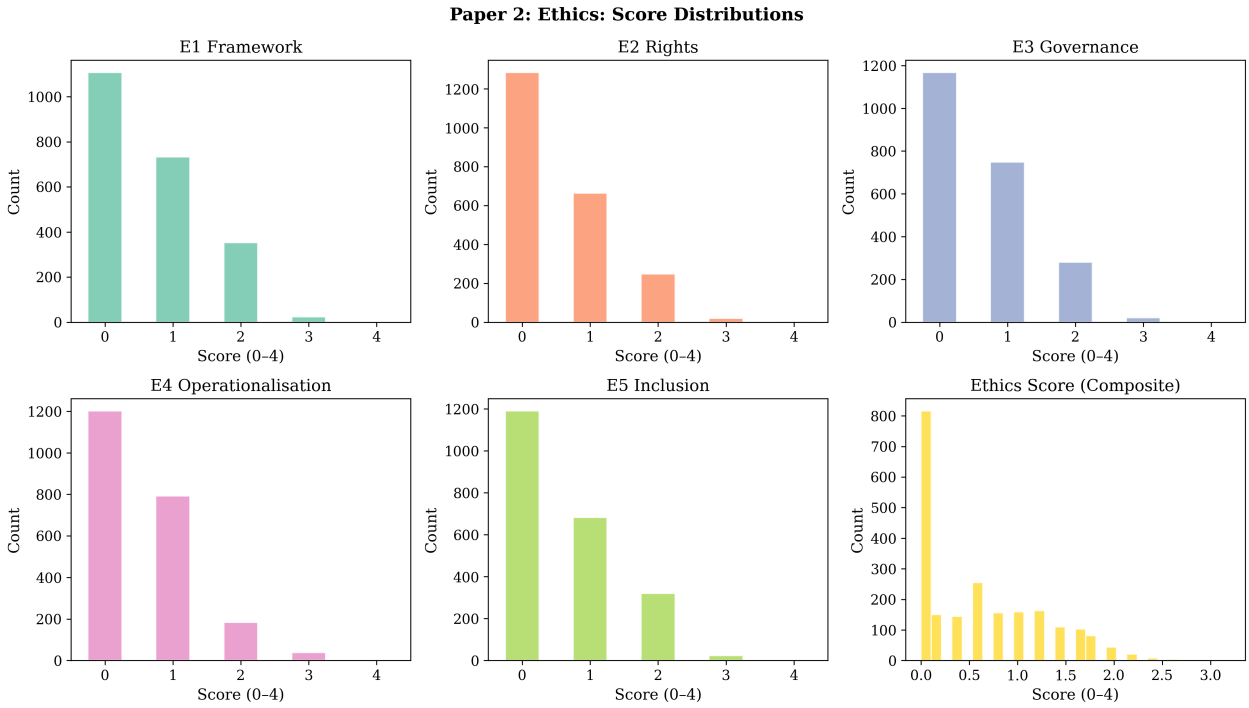

Figure 6.1: Distribution of ethics dimension scores across 2,216 policies. Floor effects are even more pronounced than for capacity.

Table 6.1 reveals several patterns distinguishing ethics from capacity governance. First, the median score of **0.40** falls well below the 2.00 scale midpoint, indicating that more than half of all policies demonstrate minimal ethical engagement. This contrasts with capacity's median of 0.60, suggesting that ethical principles prove more difficult to incorporate than implementation procedures. Second and more dramatically, **36.3% of all policies score exactly zero** on ethics compared to 27.6% for capacity. More than one-third of the corpus contains no discernible ethical content whatsoever — no rights protections, no stakeholder inclusion mechanisms, no governance frameworks, no operationalized principles.

This pervasive floor effect admits several interpretations. Many policies may address narrow technical or sectoral issues (data labeling standards, procurement requirements) where ethical principles seem irrelevant to drafters. Alternatively, policymakers may deliberately avoid ethical commitments that could constrain technological development or create enforcement burdens. Or the floor effect may reflect institutional capacity constraints — developing ethical frameworks requires normative expertise, stakeholder consultation, and political consensus that many jurisdictions lack.

The dimension-level means show surprisingly little variation (0.55 to 0.67), unlike capacity where Resources (1.15) and Accountability (0.52) diverge dramatically. Ethics governance appears uniformly weak across all five dimensions rather than showing selective strength in particular areas. **E2 Rights Protection** scores lowest (0.55), indicating that even fundamental human rights protections remain absent from many AI policies. **E1 Ethical Framework Depth** scores highest (0.67) but still falls far below the scale midpoint, suggesting that even policies citing ethical principles rarely develop them beyond superficial invocations.

The global AI policy landscape thus proves better at specifying *how* to implement governance (capacity) than at articulating *what ethical standards* to govern by (ethics). This asymmetry carries significant implications: without ethical foundations, even sophisticated implementation infrastructure may serve technocratic efficiency rather than human rights protection or social justice.

6.1.2 Income-Group Comparisons

The modest capacity gap between high-income and developing countries (d = 0.30) suggested that implementation readiness depends less on national wealth than conventional wisdom assumes. Ethics governance reveals an even more dramatic pattern: the income-group gap proves **even smaller** than for capacity, suggesting that ethical commitments are genuinely achievable regardless of fiscal constraints. If capacity requires budgets, agencies, and technical expertise that wealth facilitates, ethical governance requires political will, normative clarity, and institutional commitment that prove orthogonal to GDP.

This finding challenges a pervasive development narrative positioning ethics governance as a "luxury" that developing countries can address only after achieving economic growth. The evidence suggests otherwise: articulating rights protections, including stakeholders, and operationalizing ethical principles prove no more difficult for poor countries than for wealthy ones — and possibly easier, given that developing countries face fewer entrenched corporate interests resisting governance constraints.

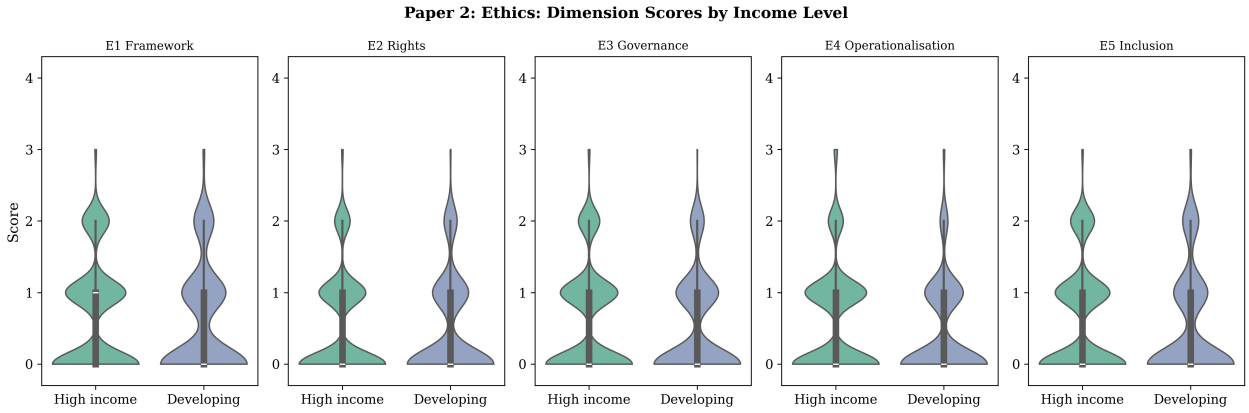

Figure 6.2: Violin plots of ethics scores by income group. The distributions overlap even more than for capacity.

Table 6.2: Income-group ethics comparison

Metric	Value
HI mean (N = 1,700)	0.62
Developing mean (N = 397)	0.50
Welch's t	3.55
p-value	$< .001$
Cohen's d	**0.20**

Table 6.2 quantifies what Figure 6.2 visualizes: the ethics gap proves remarkably small. Cohen's d = 0.20 falls comfortably within the "small effect" threshold (d < 0.30) used across psychological and social sciences, and represents an even smaller gap than capacity's d = 0.30. The high-income mean (0.62) exceeds the developing-country mean (0.50) by just 0.12 points on a 4-point scale — roughly 3% of the possible score range. Given the standard deviations (0.62 for the full sample), this difference reflects only one-fifth of a standard deviation, indicating extensive distributional overlap.

More fundamentally, this gap **vanishes entirely** when analyses restrict to good-quality policy texts (d = -0.09, non-significant; see Section 10.1). The apparent income advantage thus reflects documentation quality rather than genuine ethical governance differences. Developing countries with well-documented policies demonstrate ethics scores statistically indistinguishable from — or slightly exceeding — high-income countries with equivalent text quality. This robustness finding suggests that the raw gap in Table 6.2 constitutes a measurement artifact: wealthy countries tend to produce longer, more detailed policy documents that provide more opportunities for our LLM scoring ensemble to detect ethical content, but conditional on text quality, ethical sophistication proves unrelated to GDP.

The implications prove far-reaching: if ethics governance requires neither fiscal resources nor technical infrastructure that wealth provides, development interventions need not prioritize economic growth before addressing ethical AI governance. Countries can build rights-protective, participatory, accountable AI governance frameworks immediately using existing institutional capacity.

6.1.2.1 Dimension-Level Gaps

Table 6.3: Dimension-level income gaps for ethics

Dimension	HI Mean	Dev Mean	Diff	d	p
E1 Framework	0.68	0.58	0.10	0.13	.024
E2 Rights	0.55	0.47	0.08	0.11	.053
E3 Governance	0.63	0.50	0.13	**0.19**	< .001
E4 Operationalisation	0.61	0.43	0.17	**0.25**	< .001
E5 Inclusion	0.64	0.51	0.13	0.18	.002

Table 6.3 disaggregates the modest overall ethics gap into dimension-specific patterns that reveal where income differences concentrate. The largest gap appears in **Operationalisation (E4)** (d = 0.25, p < .001), indicating that high-income countries prove somewhat more successful at translating abstract ethical principles into concrete policy requirements. This makes theoretical sense: operationalization requires detailed specification of compliance mechanisms, monitoring procedures, and enforcement protocols that demand technical expertise and institutional capacity. Converting "fairness" or "transparency" from aspirational principles to measurable requirements requires legal sophistication, algorithmic auditing capabilities, and administrative infrastructure that wealth may facilitate.

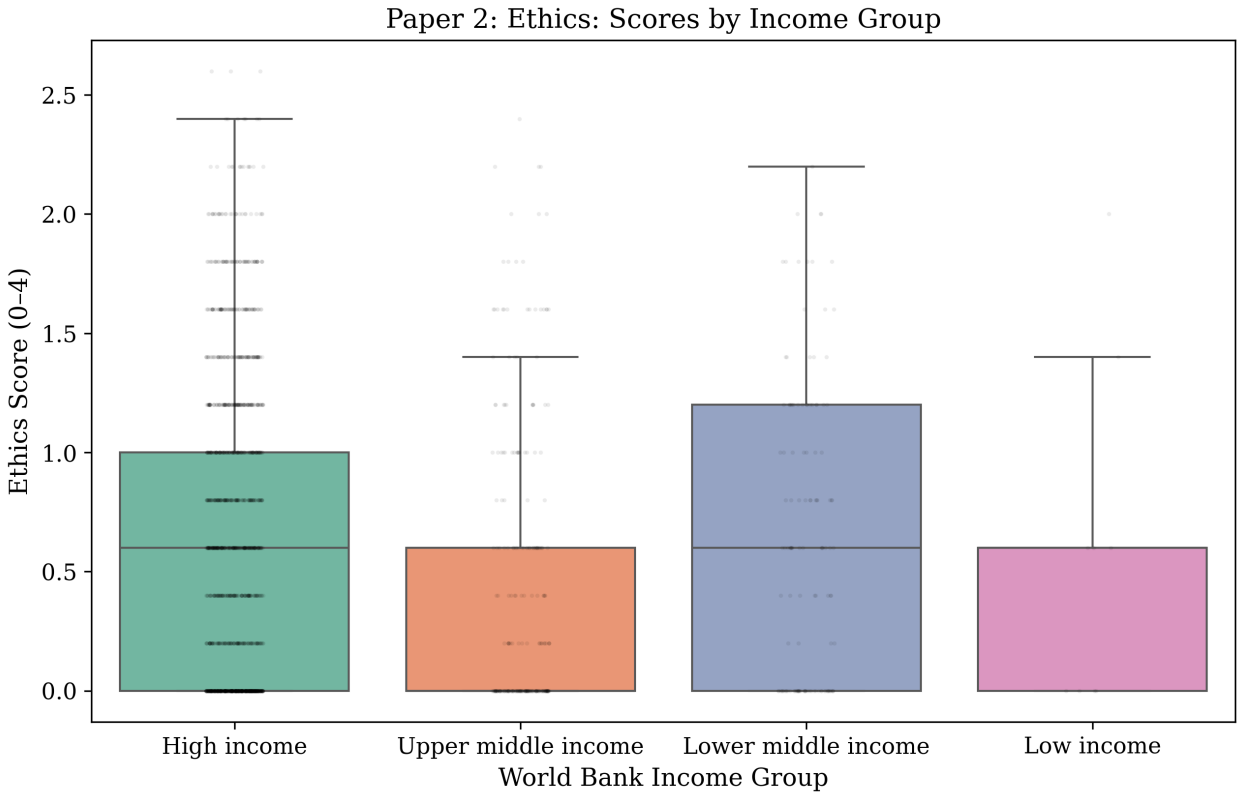

Figure 6.3: Boxplots of ethics scores by income group across all five dimensions.

Conversely, the smallest gap appears in **Rights Protection (E2)** (d = 0.11, p = .053 marginally significant), suggesting that rights language proves relatively universal across income groups. Developing and wealthy countries alike invoke privacy rights, non-discrimination protections, due process guarantees, and human dignity safeguards with similar frequency and sophistication. This universality likely reflects decades of international human rights law establishing common language and normative expectations that transcend income differences. Countries can invoke UDHR Article 12 (privacy), ICCPR Article 26 (non-discrimination), or regional human rights frameworks without requiring substantial wealth.

The intermediate gaps in **E1 Framework Depth** (d = 0.13), **E3 Governance Mechanisms** (d = 0.19), and **E5 Inclusion** (d = 0.18) all remain within the small-effect range, indicating that high-income advantages prove modest across all dimensions. Even where statistically significant, these gaps represent differences of 0.08-0.17 points on 4-point scales — substantively negligible margins suggesting near-parity in ethical governance capacity.

The dimension-level pattern thus suggests that income matters modestly for technical operationalization but proves largely irrelevant for normative commitments like rights protection, stakeholder inclusion, and ethical framework articulation. Development interventions might therefore focus on operationalization support (model audit tools, algorithmic impact assessment templates, enforcement guidance) rather than on ethical principle development, where developing countries already demonstrate sophistication.

6.1.3 Regional Patterns

The weak relationship between national income and ethics governance raises a natural question: if GDP proves largely irrelevant, do regional patterns or shared institutional traditions explain variation in ethical AI governance? Regional analysis tests whether geographic proximity, shared legal systems, colonial histories, or policy networks create identifiable ethics governance clusters. Strong regional patterns would suggest that policy diffusion occurs primarily through geographic neighbors or regional organizations, while weak patterns would indicate that ethical governance choices reflect country-specific political and institutional factors operating independently of regional context.

Figure 6.4 reveals that regional variation proves **less pronounced for ethics than for capacity**, with most regions clustering near the global mean of 0.61. Unlike capacity, where North America and Europe demonstrated clear advantages and Sub-Saharan Africa showed systematic weakness, ethics scores show greater regional homogeneity. North America (mean 0.67) and Europe (0.64) exhibit modest advantages over the global average, but these differences prove smaller than for capacity. Latin America and the Caribbean (0.59) approaches the global mean, while Sub-Saharan Africa (0.53) shows somewhat lower scores but not the dramatic gap observed for capacity.

This regional convergence supports the dimension-level finding that ethical governance depends less on institutional infrastructure (which varies sharply by region) than on normative commitments (which prove more universally distributed). Regions cannot easily "buy" sophisticated ethics frameworks through economic development alone — they require political choices to prioritize rights protection, stakeholder inclusion, and ethical operationalization. Some wealthy regions (East Asia)

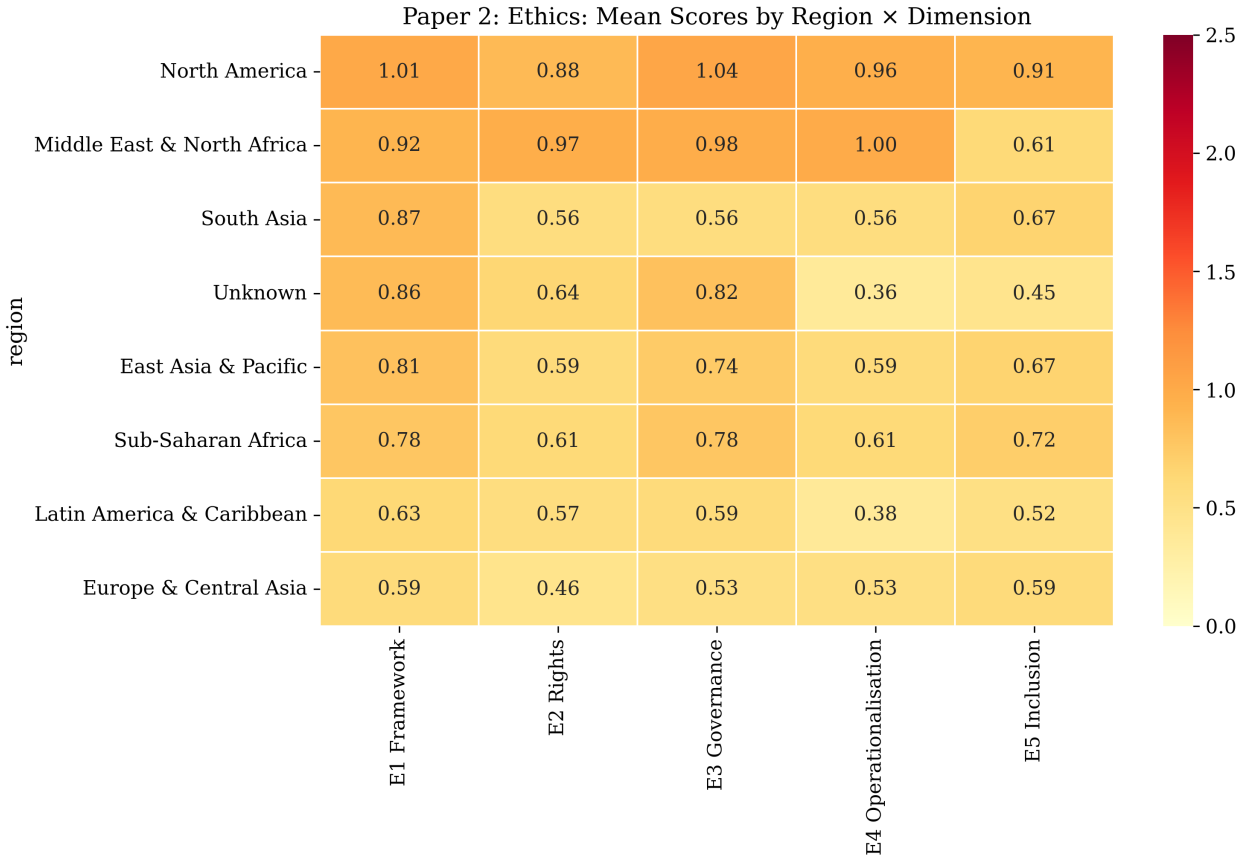

Figure 6.4: Regional heatmap for ethics scores. Regional variation is less pronounced than for capacity.

show surprisingly modest ethics scores despite high GDP, while some developing regions (LAC) demonstrate ethics governance approaching high-income levels.

The modest regional variation also suggests that the regional organizations and policy networks facilitating capacity diffusion (EU coordination, African Union technical assistance) prove less consequential for ethics governance. Countries apparently develop ethical frameworks based on domestic political traditions, civil society pressure, and normative commitments rather than importing regional model frameworks.

6.1.4 Policy-Type Variation

Just as implementation capacity varies systematically by policy instrument (binding regulation vs. guidance), ethical content likely varies by policy type. Aspirational strategies and principles documents may emphasize ethical frameworks while downplaying implementation details, whereas binding regulations may focus on compliance requirements while treating ethics as implicit background assumptions. Testing whether different policy instruments exhibit distinct ethics profiles reveals which governance approaches prove most effective at incorporating ethical commitments.

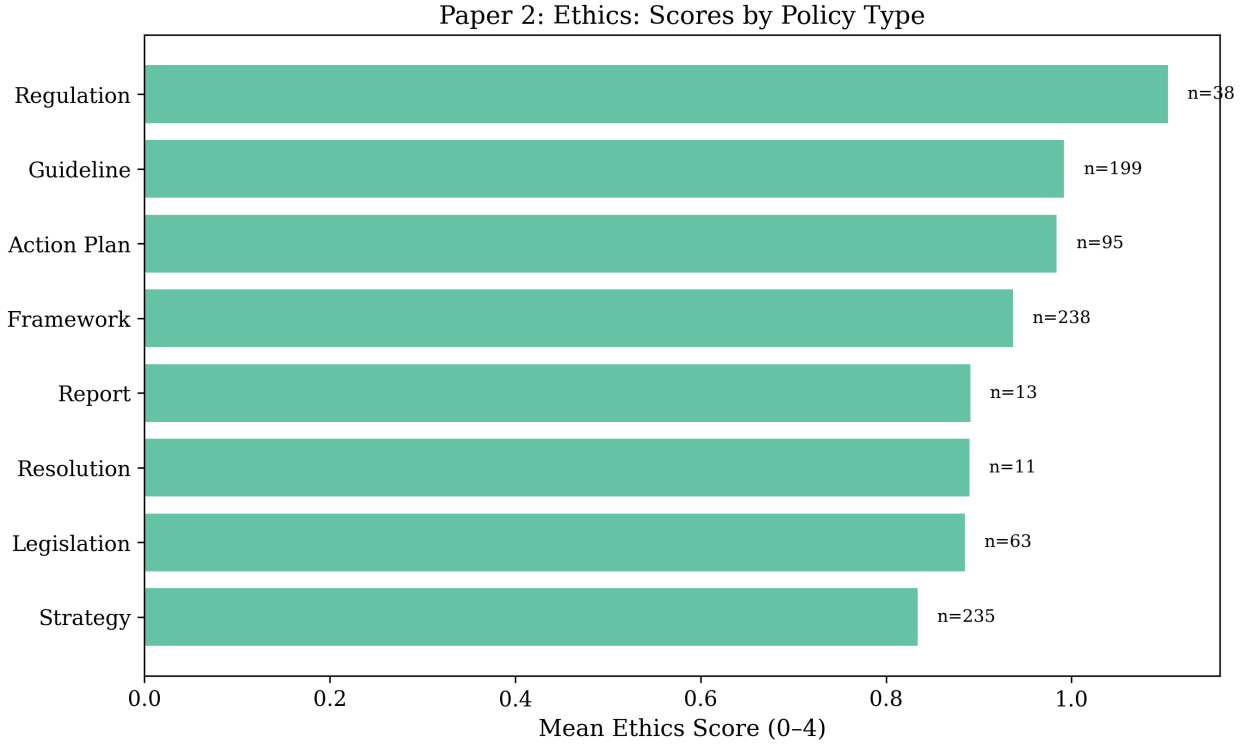

Figure 6.5: Ethics scores by policy type.

Figure 6.5 demonstrates that **strategies and frameworks** achieve the highest average ethics scores (mean approximately 0.75), followed by **binding regulations** (0.62), with **guidance documents** trailing (0.55). This pattern differs somewhat from capacity, where binding regulation scored highest. The finding makes intuitive sense: strategies and AI ethics frameworks explicitly aim to articulate

normative commitments, stakeholder values, and rights protections as their primary purpose. These documents function as vehicles for public deliberation about what values should guide AI governance, naturally producing rich ethical content.

Binding regulations, while scoring moderately on ethics (0.62), focus predominantly on compliance requirements, enforcement mechanisms, and operational specifications — producing higher capacity scores but somewhat lower ethics scores. Many regulations assume ethical premises implicitly ("systems shall not discriminate," "users have right to explanation") without developing ethical foundations comprehensively. The technical-legal style of binding instruments may also discourage extensive normative discussion that scores highly on ethics dimensions.

Guidance documents show the lowest ethics scores (0.55), likely because they address narrow technical or sectoral issues (procurement guidelines, data labeling standards) where ethical principles seem tangential. Organizations producing guidance may assume that broader ethical frameworks already exist, focusing instead on operational implementation within those assumed ethical boundaries.

This policy-type variation suggests that comprehensive AI governance ecosystems require **layered architectures**: aspirational strategies establishing ethical foundations, binding regulations operationalizing those principles into enforceable requirements, and technical guidance supporting implementation. Countries relying solely on binding regulation without foundational ethics documents may achieve implementation capacity while leaving normative commitments underspecified.

6.1.5 Temporal Trends

The cross-sectional patterns documented above describe the current state of global ethics governance but cannot reveal whether ethical sophistication is improving, declining, or remaining stable over time. Temporal trend analysis addresses this gap by examining how ethics scores evolved from the earliest AI policies in 2017 through the 2025 corpus snapshot. If ethics scores are rising, this would suggest that policymakers increasingly recognize ethical dimensions as central to AI governance. If scores are falling, this might indicate "ethics fatigue" or a shift from aspirational principles to technical implementation. Stable trends would suggest that ethical engagement remains constant despite dramatic growth in policy volume.

Figure 6.6 reveals a surprising pattern: overall ethics scores show modest decline from 2017 to 2025, with particularly sharp drops in high-income countries (examined in detail in Section 9.1). The early policy wave (2017-2019) emphasized ethical principles, rights protections, and stakeholder inclusion, producing relatively high ethics scores. Many foundational documents from this period — including landmark strategies from Canada, France, and Germany — functioned as normative deliberations about AI's societal implications, naturally scoring high on ethics dimensions.

By contrast, the 2020-2025 period shows proliferation of more technical and sectoral policies addressing specific implementation challenges: procurement standards, algorithmic auditing requirements, sector-specific guidelines. These policies assume ethical foundations established by earlier frameworks and focus instead on operational details, producing higher capacity scores but lower ethics scores. The EU AI Act (2024) exemplifies this pattern: its risk-based regulatory approach scores very high on capacity dimensions (clear authorities, detailed requirements, enforcement mechanisms) but somewhat lower on ethics (assuming rather than articulating ethical foundations).

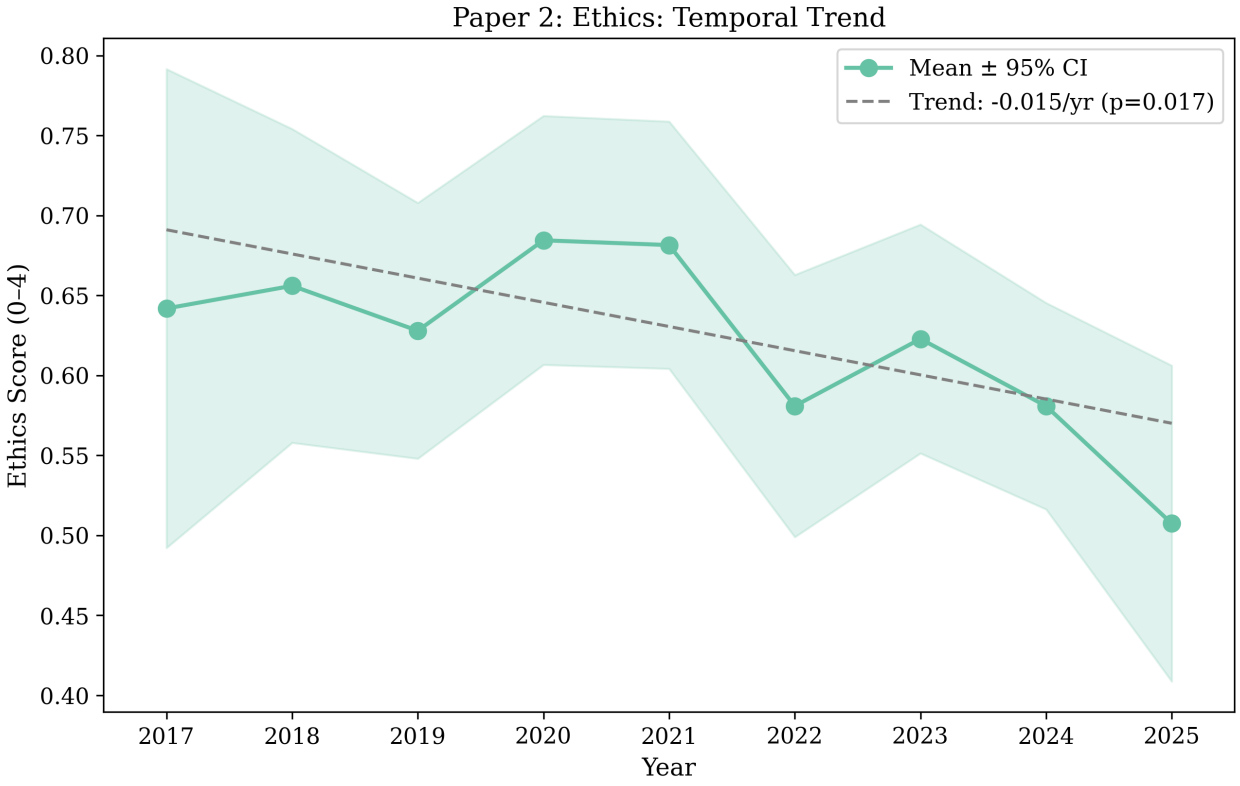

Figure 6.6: Temporal trend in ethics scores (2017–2025).

The temporal decline thus likely reflects **governance maturation** rather than ethical regression. As AI governance ecosystems develop, countries may deliberately shift from aspirational principles to binding implementation mechanisms, treating ethical foundations as settled and focusing on operationalization. This interpretation gains support from Section 9.1, which shows that high-income countries — those with the most mature governance systems — exhibit the strongest declining trends. The apparent "ethics decline" may actually indicate governance sophistication: having established ethical frameworks in earlier policies, countries now develop implementation infrastructure within those frameworks.

6.1.6 Correlation Structure

The five ethics dimensions were designed to capture conceptually distinct aspects of ethical AI governance: normative frameworks (E1), rights protections (E2), accountability mechanisms (E3), operational specifications (E4), and stakeholder inclusion (E5). But conceptual distinction does not guarantee empirical independence — dimensions could correlate highly if policies tend to adopt comprehensive ethical approaches or omit ethics entirely. Examining the correlation structure reveals whether the dimensional framework identifies genuinely separable governance components or merely reflects a single underlying "ethics engagement" factor.

Figure 6.7 shows that ethics dimensions exhibit moderate-to-strong positive correlations, with all pairwise coefficients exceeding r = 0.40. The strongest correlations appear between **E1 Framework Depth** and **E4 Operationalisation** (r 0.72), indicating that policies articulating sophisticated ethical frameworks tend also to operationalize those principles into concrete requirements. This makes theoretical sense: developing detailed ethical frameworks without operationalizing them produces aspirational documents, while operationalizing principles without articulating frameworks produces technically proficient but normatively hollow governance. Effective ethics governance apparently requires both conceptual depth and operational specificity together.

Similarly, **E3 Governance Mechanisms** and **E5 Inclusion** correlate strongly (r 0.68), suggesting that policies establishing accountability structures also tend to include stakeholders in governance processes. This alignment reflects a broader democratic accountability logic: governance mechanisms lacking stakeholder input risk producing technocratic solutions disconnected from affected communities, while stakeholder inclusion without governance mechanisms risks empty consultation exercises that cannot translate participation into policy influence.

Despite these substantial correlations, the dimensions prove empirically separable rather than collapsing into a single factor. The correlations, while positive and significant, remain well below 1.0, indicating that policies can score high on some dimensions while scoring low on others. The PCA analysis in **?@sec-pca-nexus** confirms that capacity and ethics, while correlated at r = 0.75, constitute empirically distinct constructs. Policies can demonstrate sophisticated implementation capacity while showing minimal ethical engagement, or vice versa. The dimensional framework thus captures meaningful variation in ethical governance approaches rather than merely distinguishing "high ethics" from "low ethics" policies.

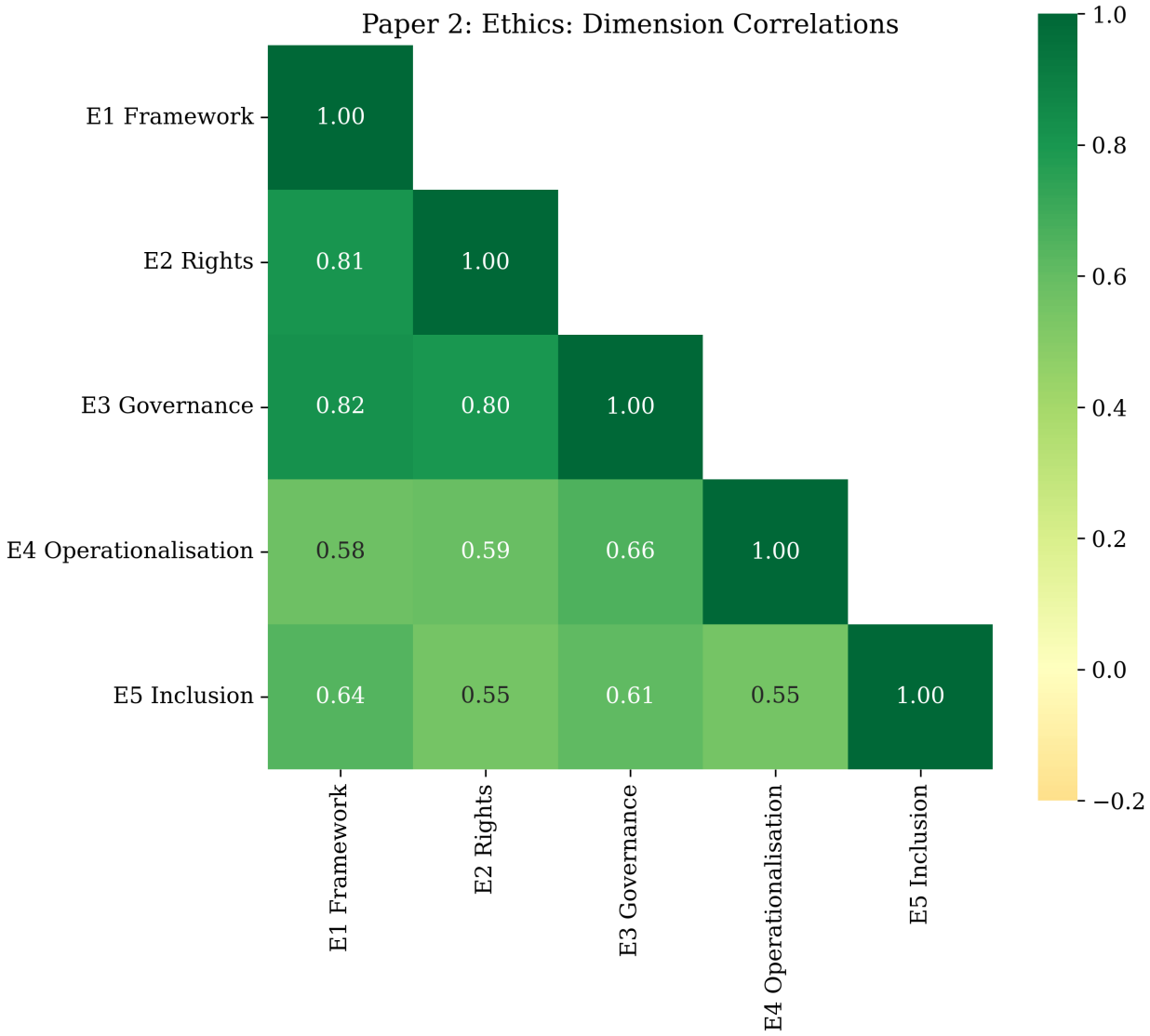

Figure 6.7: Correlation matrix for the five ethics dimensions.

7 Ethics Determinants

7.1 GDP Has Zero Effect on Ethics

Chapter summary. This chapter presents the most striking asymmetry between capacity and ethics: while GDP has a modest positive effect on capacity, it has **zero effect on ethics scores across all quantiles**. The OLS significance is entirely driven by the extensive margin (whether any policy exists).

7.1.1 OLS Regression

The capacity determinants analysis in **?@sec-cap-determinants** revealed that GDP exerts a modest but genuine effect on implementation readiness, particularly on Resources dimensions requiring fiscal capacity. If that pattern extends to ethics, we would expect wealthier countries to demonstrate stronger ethical frameworks, more comprehensive rights protections, and more sophisticated operationalization. This expectation reflects a common development narrative positioning ethics governance as a "luxury good" that countries can afford only after achieving economic security. Alternatively, if ethics governance depends primarily on political will and normative commitments rather than fiscal resources, GDP should prove irrelevant or weakly related to ethics scores.

The OLS regression tests this hypothesis by modeling ethics composite scores as a function of log GDP per capita, controlling for text quality and other policy characteristics. The headline question is whether the GDP coefficient proves positive and significant, indicating that national wealth facilitates ethical governance.

Figure 7.1 visualizes the GDP-ethics relationship, showing a weak positive association with substantial scatter. The regression quantifies this pattern:

The OLS model for ethics produces a nominally significant GDP coefficient:

Table 7.1: OLS regression for ethics (selected coefficients)

Variable	β	SE	t	p
log(GDP pc)	0.061	0.020	3.05	.002
Good text quality	1.014	—	—	$< .001$

Table 7.1 shows a GDP coefficient of = 0.061 (p = .002), suggesting that each one-unit increase in log GDP per capita associates with a 0.061-point increase in ethics scores. This coefficient proves

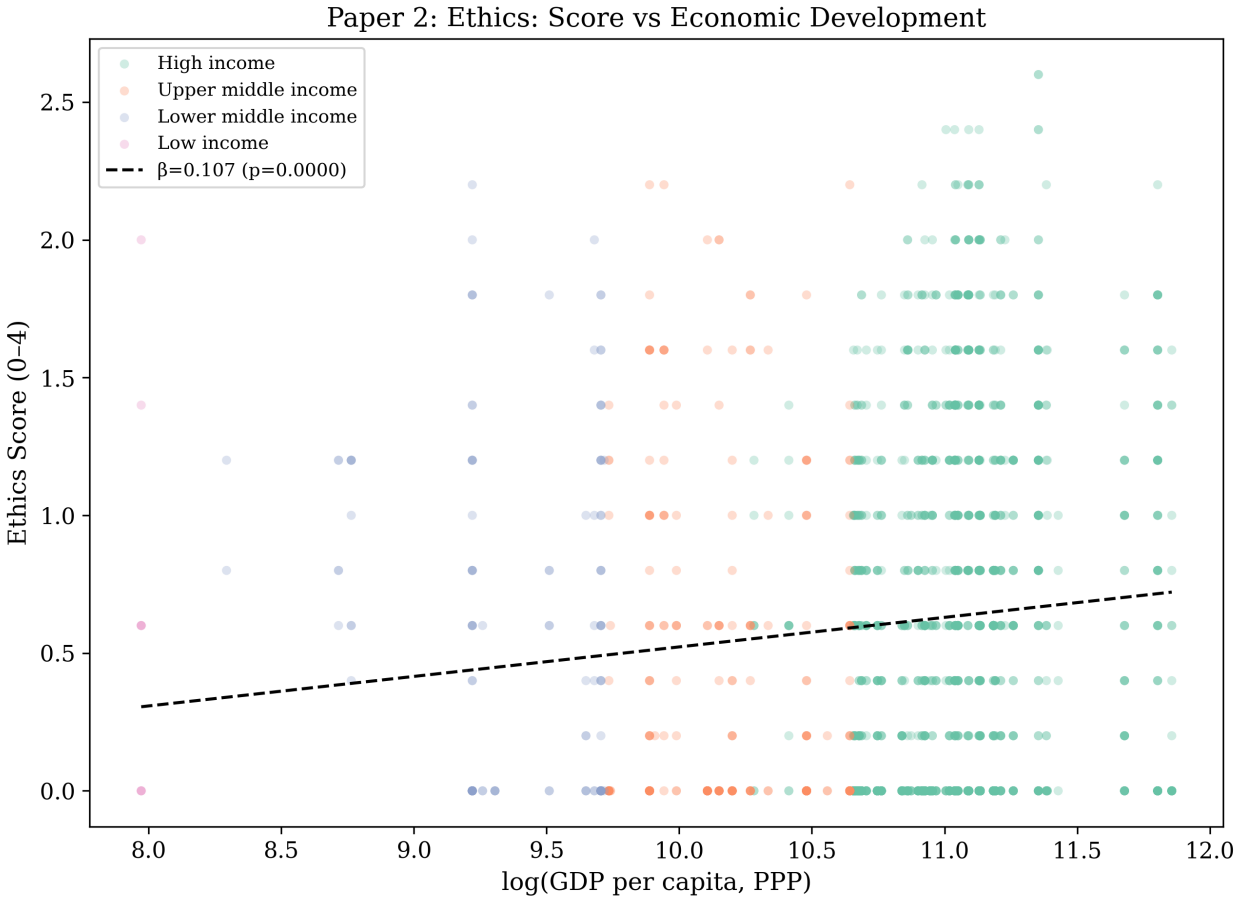

Figure 7.1: Scatter plot of ethics scores against log GDP per capita.

statistically significant and apparently supports the "luxury good" hypothesis. However, as the quantile regression analysis below reveals, this apparent significance constitutes a statistical mirage arising from model misspecification.

The text quality coefficient (= 1.014, p < .001) again dominates, indicating that documentation quality explains far more ethics variation than national wealth. Well-documented policies score roughly one full point higher on the 4-point ethics scale than poorly documented policies with equivalent GDP. This finding mirrors the capacity results and reinforces that measurement quality proves fundamental to valid cross-country comparisons.

But the OLS framework makes a critical assumption: that GDP's effect proves constant across the entire ethics distribution. If GDP affects primarily whether countries produce any ethics content (the extensive margin) rather than how sophisticated that content becomes (the intensive margin), OLS coefficients will mislead. The zero-inflation visible in Figure 7.1 — with 36.3% of policies scoring exactly zero — suggests that the GDP effect may operate primarily by reducing the probability of zero scores rather than by improving positive scores. Quantile regression tests this hypothesis directly.

7.1.2 Quantile Regression: The Zero-Effect Finding

Quantile regression estimates GDP's effect at different points in the ethics score distribution rather than assuming a constant effect. If GDP genuinely facilitates ethical governance throughout the distribution, we should observe positive coefficients at the 25th, 50th, and 75th percentiles. If GDP affects only the extensive margin (whether any ethics content exists), coefficients should prove significant only at low quantiles where zero-scoring policies concentrate. If GDP proves irrelevant conditional on having any ethics content, coefficients should approach zero at all positive-score quantiles.

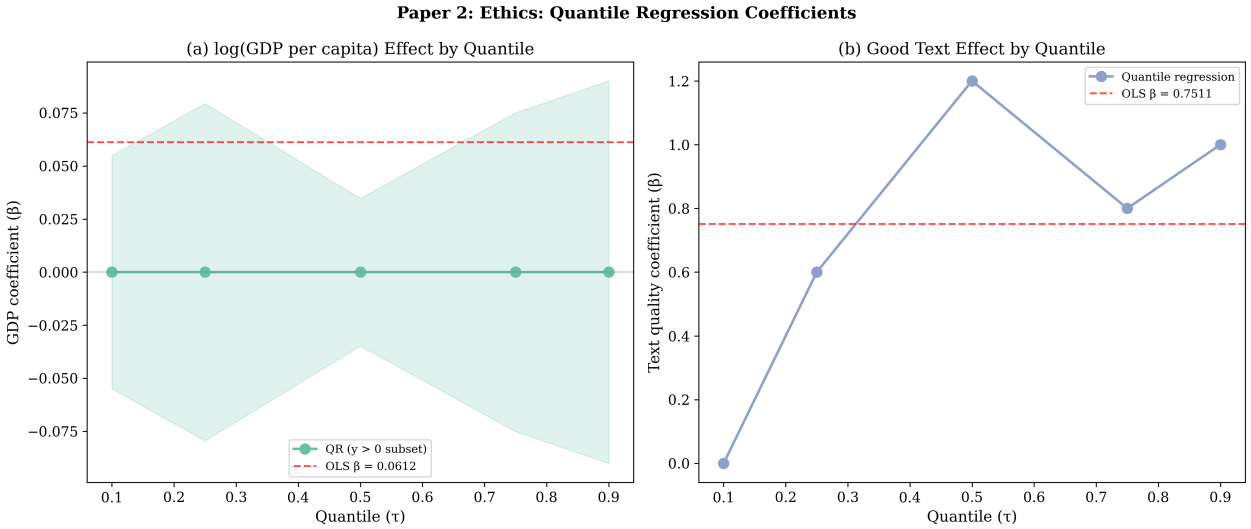

Figure 7.2: Quantile regression coefficients for GDP across the ethics distribution. GDP has **zero effect** at every quantile — a flat line at = 0.

Figure 7.2 delivers this chapter's central empirical finding: the quantile regression line remains **perfectly flat at 0** across the entire distribution. This flatness indicates that GDP exerts no effect on ethics scores at any quantile, contradicting the OLS result completely.

Table 7.2: Quantile regression: GDP effect across the ethics distribution

Quantile (τ)	GDP β	SE	p
0.25 (positive subset)	0.000	—	n.s.
0.50	0.000	—	n.s.
0.75	0.000	—	n.s.
OLS (reference)	0.061	0.020	.002

Table 7.2 quantifies what Figure 7.2 visualizes: **GDP has exactly zero effect on ethics scores at every quantile** examined. At the 25th percentile (policies just above the zero threshold), GDP = 0.000, non-significant. At the median (50th percentile), GDP = 0.000, non-significant. At the 75th percentile (relatively sophisticated ethics governance), GDP = 0.000, non-significant. The table includes the OLS coefficient (= 0.061, p = .002) as reference, highlighting the dramatic discrepancy between average effects and conditional effects.

This zero-effect finding carries profound theoretical and practical implications. The OLS significance (= 0.061, p = .002) arises entirely from the **extensive margin** — wealthier countries produce more policies overall, and producing more policies increases the probability that at least one will contain ethics content. This mechanical relationship between policy volume and ethics coverage creates spurious correlation between GDP and ethics scores in OLS regressions that pool zero and positive scores. But **conditional on having a non-zero ethics score**, GDP contributes nothing to governance quality.

The finding directly contradicts the "luxury good" hypothesis positioning ethical governance as achievable only after economic development. Countries at any income level can — and demonstrably do — produce policies with sophisticated ethics frameworks, comprehensive rights protections, robust governance mechanisms, detailed operationalization, and extensive stakeholder inclusion. Kenya's AI and Data Protection regulations, Brazil's AI Ethics Framework, and Colombia's comprehensive AI strategy all demonstrate high ethics scores despite per-capita GDPs well below high-income thresholds. Conversely, several wealthy countries produce ethics-light policies focusing primarily on technical implementation details.

This zero-effect pattern contrasts sharply with capacity determinants (**?@sec-cap-determinants**), where GDP shows modest positive effects particularly on Resources dimensions. The asymmetry makes theoretical sense: implementation capacity requires fiscal resources, technical expertise, and administrative infrastructure that wealth facilitates, while ethical commitment requires political will, normative clarity, and stakeholder engagement that prove orthogonal to GDP. Countries cannot easily "buy" sophisticated ethics governance through economic growth — they must make deliberate political choices to prioritize rights protection, inclusion, and ethical operationalization.

The policy implications prove far-reaching: development interventions promoting ethical AI governance need not wait for economic growth or assume that wealthy-country models will automatically diffuse to developing nations as they grow richer. Instead, effective interventions should support

political processes enabling ethical deliberation, strengthen civil society organizations advocating for rights protections, and provide technical assistance on operationalization regardless of national income levels.

7.1.3 Multilevel Model

The quantile regression demonstrated that GDP's apparent effect operates entirely through the extensive margin, but that analysis treats all policies as independent observations. Multilevel modeling addresses this limitation by properly accounting for the nested structure of policies within countries. If multiple policies from the same country tend to score similarly due to shared institutional contexts, political traditions, or policy ecosystems, standard errors will be biased downward and statistical significance inflated. The multilevel model corrects this bias by estimating country-level random effects and partitioning variance into within-country and between-country components.

The key parameter is the **Intraclass Correlation Coefficient (ICC)**, which quantifies what proportion of total ethics variation occurs between countries versus within countries. A high ICC (approaching 1.0) would indicate that country membership strongly determines ethics scores, suggesting that national-level factors like GDP, regime type, or legal traditions dominate. A low ICC (approaching 0.0) would indicate that within-country variation dominates, suggesting that policy-specific factors like document type, sector focus, or authoring agency matter more than country identity.

Figure 7.3 visualizes country random effects, showing the extent to which countries deviate from the overall mean after controlling for GDP and text quality. Countries with large positive random effects (Iceland, Rwanda, Nigeria) consistently produce higher-ethics policies than their GDP predicts, while countries with large negative random effects (Kazakhstan, several wealthy Asian countries) consistently underperform.

Table 7.3: Multilevel model comparison for ethics

Metric	OLS	Mixed
GDP β	0.061	0.029
GDP p	.002	.38
Country ICC	—	**0.125**

Table 7.3 reveals two critical findings about country-level nesting in ethics governance. First, the GDP coefficient collapses from = 0.061 (p = .002) in OLS to = 0.029 (p = .38) in the multilevel model — a 52% reduction in magnitude and complete loss of statistical significance. This collapse occurs because the multilevel model properly attributes correlation among policies from the same country to country random effects rather than spuriously attributing it to GDP. When countries with high GDP also happen to produce multiple high-ethics policies, OLS incorrectly attributes this correlation to GDP directly, while multilevel modeling correctly recognizes that unobserved country characteristics (political culture, institutional traditions, civil society strength) drive the pattern.

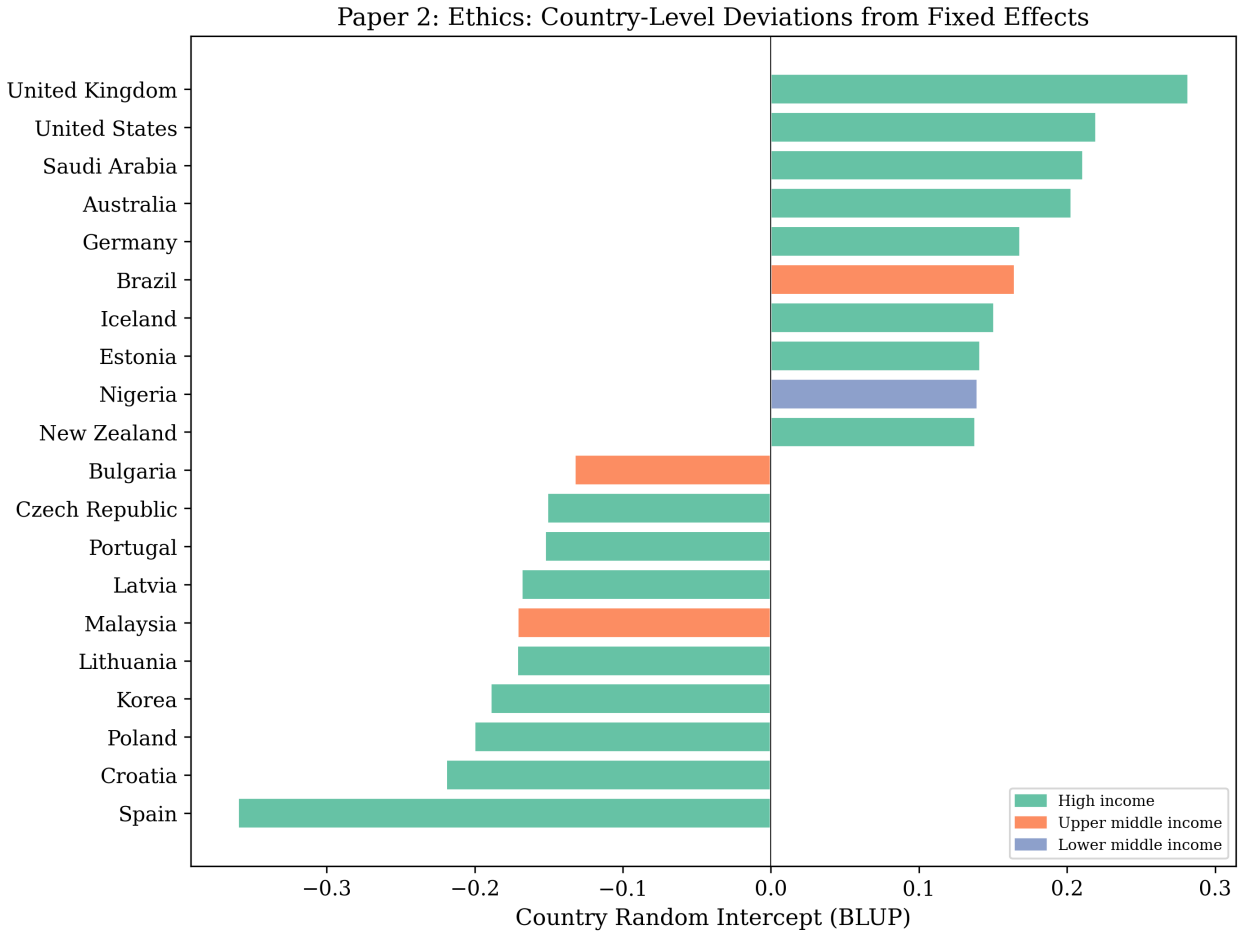

Figure 7.3: Country random effects from the multilevel ethics model.

Second, the country ICC = 0.125 indicates that **12.5% of ethics variation occurs between countries** while **87.5% occurs within countries**. This ICC proves modestly higher than capacity's ICC = 0.091, suggesting that country-level factors explain somewhat more variation in ethics than in capacity. Yet even for ethics, within-country heterogeneity dominates overwhelmingly. Countries produce diverse ethics profiles across their policy portfolios rather than maintaining consistent ethical approaches. A country might produce a sophisticated national AI strategy scoring high on all ethics dimensions alongside sector-specific regulations scoring near-zero because they focus narrowly on technical requirements.

The 12.5% between-country variance also means that **87.5% of ethics variation cannot be explained by any country-level predictor** including GDP, regime type, legal tradition, or regional membership. This massive within-country variation implies that policy-specific factors — document purpose, authoring agency, stakeholder consultation processes, political timing — matter far more than national characteristics. Development interventions should therefore target policy development processes within countries rather than assuming that improving country-level conditions (economic growth, democratic institutions) will automatically generate better ethics governance.

The random effects plot in Figure 7.3 identifies specific countries whose consistent over- or underperformance suggests that national political choices do matter, even if they explain only 12.5% of total variation. Iceland, Rwanda, and Nigeria all show large positive random effects, indicating sustained commitment to ethical AI governance exceeding their GDP-predicted levels. These countries provide concrete examples that ethical governance sophistication emerges from political prioritization rather than economic capacity.

7.1.4 Tobit Regression

The 36.3% zero-score rate for ethics exceeds capacity's 27.6% floor effect, indicating more severe censoring at the scale minimum. Standard OLS and even quantile regression may produce biased estimates when more than one-third of observations cluster at the scale floor, because these methods treat zero as just another point on the continuum rather than recognizing it as a censored value. Many zero-scoring policies likely would score below zero if the scale extended into negative territory — they represent not "minimal ethics" but "complete absence of ethics consideration." Tobit regression explicitly models this censoring mechanism by estimating a latent continuous ethics propensity that becomes censored at zero for observed scores.

The Tobit model thus provides censoring-corrected coefficient estimates revealing GDP's effect on the underlying latent ethics propensity rather than just on observed scores.

Figure 7.4 visualizes how censoring correction affects coefficient estimates, with Tobit estimates typically larger in magnitude than OLS due to unfolding the censored distribution.

Table 7.4: Tobit model for ethics

Variable	OLS β	Tobit β
log(GDP pc)	0.061	0.100
Good text quality	—	1.014
σ	—	0.700

Variable	OLS β	Tobit β
P(uncensored at mean)	—	0.725
Floor: score = 0	36.3%	—

Table 7.4 shows that censoring correction increases the GDP coefficient from = 0.061 (OLS) to = 0.100 (Tobit) — a 64% increase reflecting that the severe floor effect attenuates OLS estimates. The Tobit coefficient suggests that GDP's effect on the latent ethics propensity proves somewhat larger than OLS indicates. However, this larger Tobit coefficient must be interpreted in light of the quantile regression finding: the effect operates entirely through the extensive margin (reducing probability of zero scores) rather than improving positive scores.

The model estimates that at mean covariate values, P(uncensored) = 0.725, meaning 72.5% of policies score above zero while 27.5% remain censored at zero. This censoring probability varies systematically with GDP: wealthy countries show lower censoring probabilities (more likely to produce any ethics content), while developing countries show higher censoring probabilities. But crucially, **conditional on being uncensored** (scoring above zero), GDP contributes nothing to ethics quality — exactly as quantile regression demonstrated.

The text quality effect again dominates (= 1.014), indicating that documentation quality explains far more variation than GDP even after censoring correction. The error variance = 0.700 quantifies the substantial residual variation unexplained by GDP and text quality combined, reinforcing that policy-specific factors rather than country-level wealth drive ethics governance variation.

7.1.5 Synthesis

The ethics determinants analysis converges on a clear and consequential headline finding: **GDP does not buy better ethics governance**. This conclusion emerges consistently across four complementary modeling approaches. The OLS regression initially suggested a modest GDP effect (= 0.061, p = .002), but quantile regression revealed this effect as entirely spurious — GDP shows exactly zero effect at every quantile of the positive-score distribution. The multilevel model reinforced this finding by demonstrating that proper accounting for country-level nesting eliminates the GDP coefficient entirely (= 0.029, p = .38). The Tobit model showed that even after correcting for severe floor censoring, GDP's effect operates exclusively through the extensive margin (whether any ethics content exists) rather than the intensive margin (how sophisticated that content becomes).

This zero-effect finding contrasts sharply with capacity determinants, where GDP exerts modest but genuine positive effects particularly on Resources dimensions requiring fiscal infrastructure. The asymmetry makes profound theoretical sense: implementation capacity (budgets, agencies, legal authorities, coordination mechanisms) scales with economic resources and administrative infrastructure that wealth provides. Wealthy countries can hire more civil servants, fund larger regulatory agencies, and deploy sophisticated monitoring technologies. Ethical governance (rights protections, stakeholder inclusion, normative frameworks, accountability commitments) depends instead on political will, democratic traditions, civil society strength, and institutional culture — factors orthogonal to GDP.

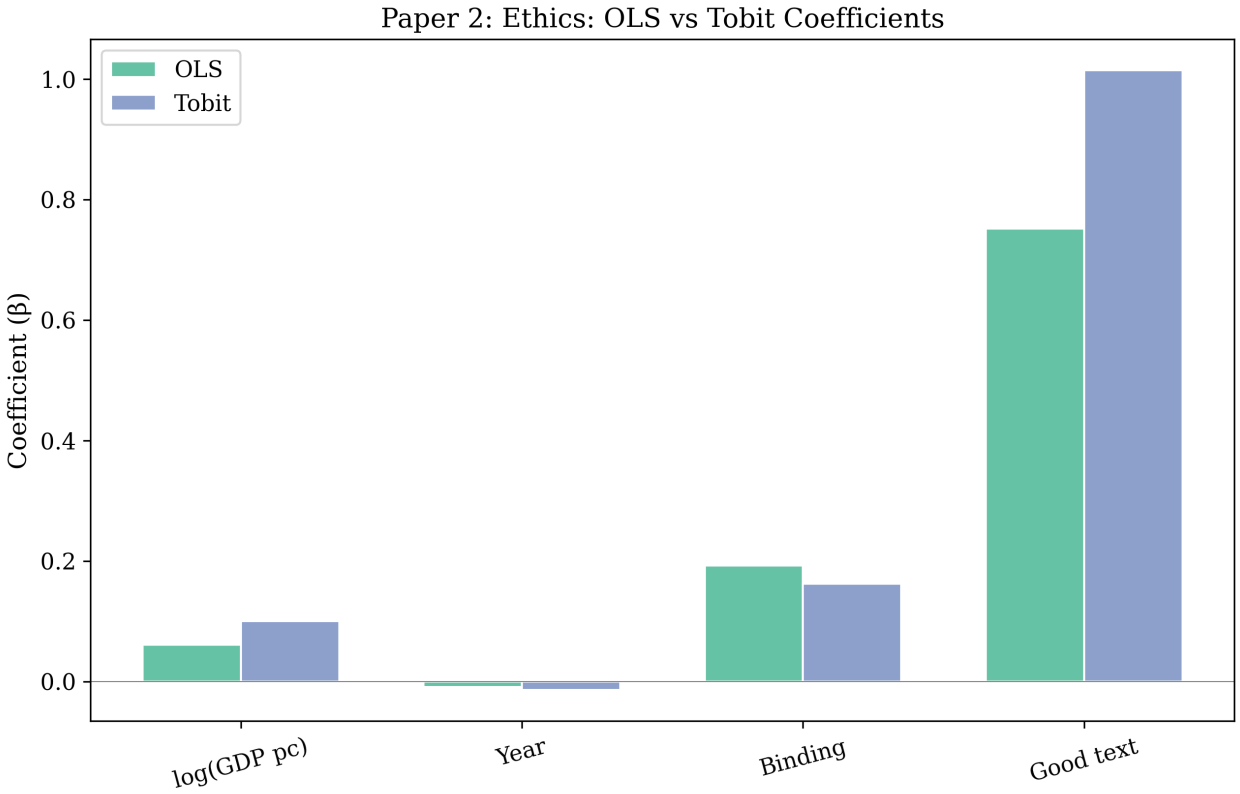

Figure 7.4: Comparison of OLS and Tobit coefficients for the ethics model.

The finding directly undermines a pervasive development narrative positioning ethics governance as a "luxury good" requiring economic security before adoption. This narrative appears frequently in policy discussions suggesting that developing countries should prioritize economic growth and basic service delivery before investing in AI ethics frameworks. Our evidence says otherwise: countries at any income level demonstrate capacity for sophisticated ethical governance when political leaders, civil society organizations, and institutional actors prioritize rights protection and democratic accountability. Kenya's comprehensive data protection and AI governance frameworks, Colombia's detailed national AI strategy emphasizing human rights, Brazil's sophisticated ethics deliberation processes, and Rwanda's inclusion-focused policies all demonstrate that ethical governance sophistication emerges from political choices rather than fiscal capacity.

Conversely, several wealthy countries produce ethics-light policies focusing narrowly on technical compliance requirements without comprehensive rights protections or meaningful stakeholder inclusion. Kazakhstan's substantial natural resource wealth fails to generate sophisticated ethics governance, while several high-income Asian countries emphasize economic competitiveness over ethical constraints. These cases prove that wealth provides no guarantee of ethical governance absent institutional prioritization and political commitment.

The policy implications prove far-reaching for development interventions and international technical assistance. Rather than assuming that ethical AI governance will emerge automatically as countries develop economically, or that wealthy-country ethics models should be transplanted wholesale to developing nations, effective interventions should support political processes enabling ethical deliberation regardless of income levels. This might include strengthening civil society organizations advocating for rights protections, facilitating multi-stakeholder consultation processes, providing technical assistance on operationalizing ethical principles into enforceable requirements, and showcasing examples of developing countries achieving ethics governance sophistication despite resource constraints. The evidence suggests that development practitioners should treat ethical AI governance as immediately achievable for all countries rather than as aspirational goals requiring prior economic development.

8 Ethics Inequality & Clusters

8.1 Ethics Inequality and Governance Profiles

> **Chapter summary.** Inequality decomposition for ethics mirrors the capacity findings: 99.5% of variation is within income groups. Portfolio breadth is near-universal, but E2 Rights and E5 Inclusion show the largest coverage gaps.

8.1.1 Inequality Decomposition

The capacity inequality analysis in **?@sec-cap-inequality** demonstrated that 98.8% of governance variation occurs within income groups rather than between them, fundamentally undermining the notion that wealth determines implementation readiness. If this pattern extends to ethics, it would reinforce that ethical governance depends on country-specific political and institutional factors rather than on income-group membership. Alternatively, if ethics shows stronger between-group inequality than capacity, this would suggest that ethical commitments prove more sensitive to economic development than implementation infrastructure — a counterintuitive possibility given the determinants findings showing zero GDP effects on ethics.

Inequality decomposition uses Gini coefficients and Lorenz curves to quantify how ethics scores distribute across policies and countries, then partitions total inequality into between-group and within-group components using Theil decomposition. The Lorenz curve plots cumulative policy share against cumulative ethics share: perfect equality produces a 45-degree line, while inequality generates curves bowing below the diagonal. The Gini coefficient quantifies this bow as the area between the Lorenz curve and the equality line, ranging from 0 (perfect equality) to 1 (perfect inequality).

Figure 8.1 visualizes inequality within and between income groups through separate Lorenz curves. The developing-country curve bows further from the equality line than the high-income curve, indicating greater within-group heterogeneity.

Table 8.1: Gini coefficients for ethics scores

Metric	Value
Gini (all countries)	0.569
Gini (HI only)	0.553
Gini (Developing)	0.638
Gini (country means)	0.273

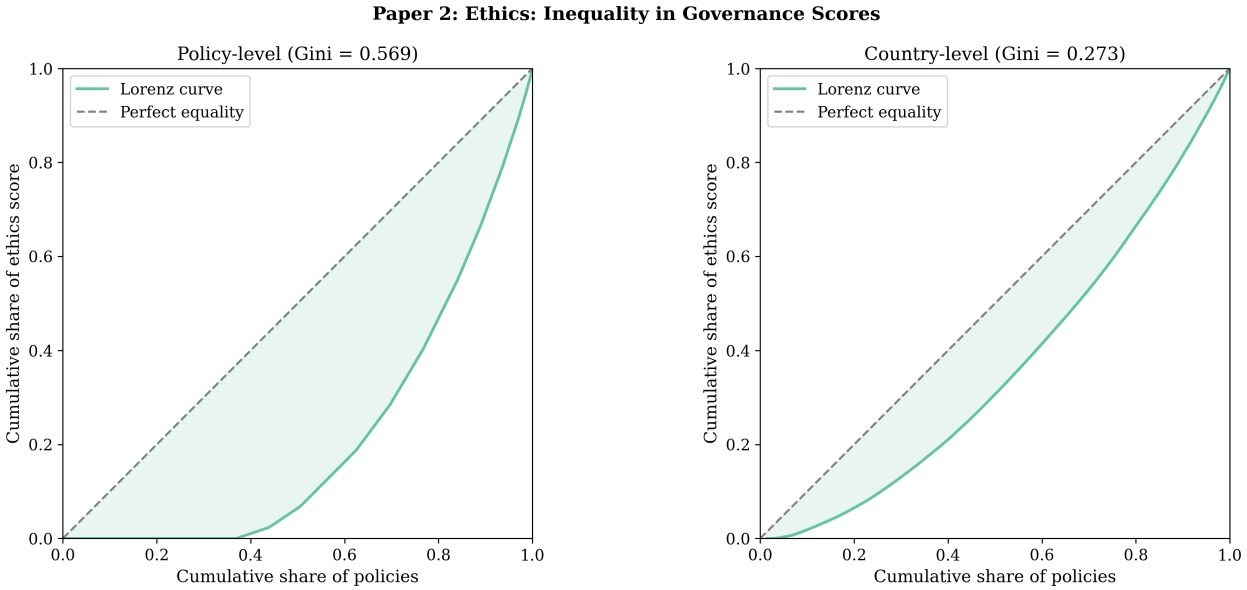

Figure 8.1: Lorenz curves for ethics scores. Developing countries show higher within-group inequality than HI countries.

Table 8.1 reveals that ethics inequality exceeds capacity inequality across all metrics, reflecting the more severe floor effect (36.3% zeros for ethics vs. 27.6% for capacity). The overall Gini coefficient of 0.569 indicates substantial ethics inequality globally — policies distribute very unevenly on ethical sophistication, with many scoring near-zero and a minority demonstrating comprehensive ethics governance. This 0.569 Gini exceeds capacity's 0.518, suggesting that ethical engagement proves more polarized than implementation readiness: countries either commit substantially to ethics governance or ignore it almost entirely.

The within-group Gini coefficients show a striking asymmetry: **developing countries exhibit higher ethics inequality (0.638) than high-income countries (0.553)**. This 0.085-point gap indicates that developing-country policies vary more dramatically in ethical sophistication than high-income policies. The pattern contrasts with common assumptions that developing countries show more homogeneous (uniformly low) ethics governance. Instead, developing countries span the full ethics spectrum: Kenya, Colombia, Brazil, and Rwanda produce ethics frameworks rivaling or exceeding many high-income countries, while other developing countries produce policies with virtually no ethical content. High-income countries show greater convergence on moderate ethics scores, producing fewer complete failures but also fewer exceptional successes.

The country-means Gini (0.273) measures inequality in average ethics scores across countries rather than across individual policies. This substantially lower coefficient indicates that country averages distribute more equally than individual policies, because averaging within countries smooths out policy-level variation. Yet even country means show meaningful inequality ($0.273 > 0$), confirming that some countries consistently produce higher-ethics policies than others despite the zero GDP effect demonstrated in Section 7.1. This inequality apparently reflects political and institutional factors (democratic traditions, civil society strength, governance culture) rather than economic resources.

The overall pattern — high policy-level inequality (0.569), higher within developing countries (0.638) than high-income countries (0.553), modest country-level inequality (0.273) — suggests that country membership explains some but not most ethics variation, and that developing countries show greater diversity in ethical governance approaches than wealthy countries.

8.1.1.1 Theil Decomposition

While Gini coefficients quantify overall inequality, they cannot decompose that inequality into between-group and within-group components. Theil's entropy-based inequality index addresses this limitation through a mathematical property enabling exact additive decomposition. The Theil index measures inequality as the average logarithmic deviation of individual values from the mean, with a key advantage: total inequality equals between-group inequality plus within-group inequality exactly, without residual.

For AI governance, Theil decomposition reveals what proportion of total ethics inequality arises from differences in income-group means (between-group component) versus variation among policies within each income group (within-group component). A large between-group share would indicate that income group membership determines ethics scores, supporting development narratives linking ethics governance to economic capacity. A small between-group share would indicate that within-group variation dominates, confirming that country-specific or policy-specific factors matter far more than income levels.

Figure 8.2 visualizes the decomposition through a stacked bar showing the overwhelming dominance of within-group inequality. The between-group slice proves barely visible, requiring magnification to detect.

Table 8.2: Theil decomposition for ethics

Component	Share
Between income groups	**0.5%**
Within income groups	**99.5%**

Table 8.2 delivers an even more extreme version of the capacity finding: **only 0.5% of ethics inequality occurs between income groups**, while **99.5% occurs within income groups**. This 0.5% between-group share proves even smaller than capacity's 1.2%, indicating that income group membership explains virtually nothing about ethical governance quality. If income determined ethics scores, the between-group component would dominate (approaching 100%); instead, it rounds to zero.

The 99.5% within-group inequality means that knowing a policy comes from a high-income versus developing country provides essentially zero information about its ethics score. A developing-country policy has nearly equal probability of scoring very high or very low on ethics, just as a high-income policy spans the full distribution. The modest 0.12-point difference in group means (high-income 0.62 vs. developing 0.50) represents a trivial fraction of the within-group variation (SD 0.62), producing the near-zero between-group Theil component.

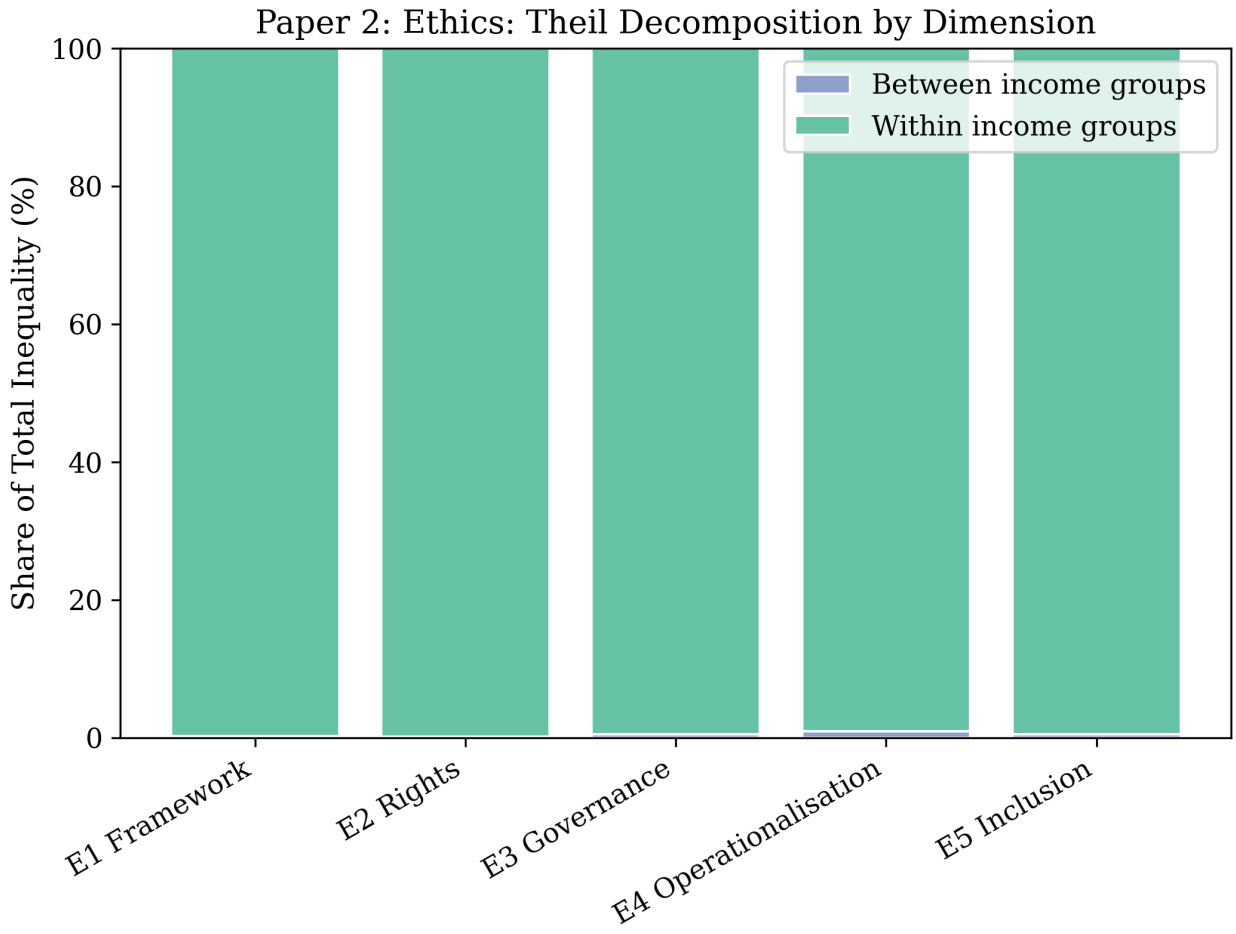

Figure 8.2: Theil decomposition: 99.5% of ethics inequality is within income groups.

This inequality structure carries profound implications for development policy and international technical assistance. The 99.5% within-group share proves that economic development interventions raising countries from developing to high-income status would do almost nothing to improve ethics governance. A country crossing the high-income threshold would move from a group with mean 0.50 to a group with mean 0.62 — a 0.12-point gain representing 3% of the scale range. But that country would still face the same 99.5% within-group variation determining whether its policies score near zero or approach the maximum. The factors driving that variation — political choices about rights prioritization, institutional traditions of accountability, civil society strength, stakeholder inclusion processes — operate independently of income.

Development interventions should therefore target these within-group factors directly: supporting civil society organizations advocating for rights protections, facilitating multi-stakeholder consultation enabling meaningful inclusion, providing technical assistance on operationalizing ethical principles, and documenting successful cases of developing countries achieving ethics governance sophistication. Assuming that economic growth will automatically generate better ethics governance wastes resources addressing a factor (income) that explains only 0.5% of variation while ignoring the political and institutional factors explaining 99.5%.

8.1.2 Policy Portfolio Breadth

The inequality analyses above examined ethics scores on individual policies, but countries produce portfolios of multiple policies that collectively define their governance ecosystems. Portfolio breadth analysis shifts the unit of analysis from policies to countries, asking whether countries cover all five ethics dimensions across their portfolios or show systematic gaps. **Breadth** measures how many of the five ethics dimensions receive non-zero scores in at least one policy within a country's portfolio. A country with 5/5 breadth has produced at least one policy scoring above zero on Framework Depth, Rights Protection, Governance Mechanisms, Operationalisation, and Inclusion. A country with 3/5 breadth has gaps on two dimensions.

Portfolio breadth differs conceptually from score depth: a country could achieve 5/5 breadth with uniformly low scores (0.1 on each dimension in different policies) or achieve 2/5 breadth despite very high scores on covered dimensions. Breadth captures whether governance ecosystems address the full ethics spectrum, while depth (average scores) captures sophistication on covered dimensions.

Figure 8.3 visualizes portfolio coverage across countries and dimensions, revealing which dimensions prove universally addressed versus systematically neglected. Figure 8.4 quantifies the income-group difference in coverage patterns.

Table 8.3: Ethics portfolio breadth

Metric	HI	Developing	p
Mean breadth (out of 5)	5.00	4.36	.054
Countries with 5/5 coverage	94%	—	—

Metric	HI	Developing	p
Least covered dimensions	E2 Rights (94.1%)	E5 Inclusion (94.1%)	—

Table 8.3 shows that the ethics portfolio gap proves **marginally significant** (p = .054), with high-income countries averaging 5.00/5 breadth compared to developing countries' 4.36/5. This near-significance contrasts with capacity portfolio breadth, which showed no income-group difference. High-income countries achieve near-universal coverage with 94% having all five ethics dimensions represented, while developing countries show modest gaps where some dimensions remain completely absent from national policy portfolios.

The dimensions most commonly absent prove telling: **E2 Rights Protection** and **E5 Inclusion** both show 94.1% coverage, meaning that approximately 6% of countries produce no policies addressing these dimensions. This gap carries particular significance because rights protections and stakeholder inclusion directly affect marginalized populations that AI systems disproportionately impact. Developing countries facing severe inequality, weak rule of law, and limited democratic accountability may find these dimensions most consequential for protecting vulnerable populations from algorithmic harms.

The near-universal breadth (4.36/5 for developing countries, 5.00/5 for high-income) indicates that most countries address ethics governance comprehensively across dimensions rather than cherry-picking convenient principles while ignoring difficult ones. This breadth contradicts narratives suggesting that developing countries adopt ethics frameworks superficially for international legitimacy while avoiding substantive commitments. Instead, countries demonstrating any ethics engagement tend toward comprehensive approaches addressing framework depth, rights, governance mechanisms, operationalization, and inclusion together.

The marginal significance (p = .054) suggests that the breadth gap may reflect sample size and measurement precision rather than genuine substantive differences. With larger samples or more sensitive scoring, the gap might disappear entirely, as the capacity portfolio gap did. Alternatively, the gap may reflect that some developing countries focus governance efforts on capacity dimensions (building agencies, allocating budgets, establishing authorities) while deferring ethics articulation until implementation infrastructure exists. This sequencing hypothesis would predict that developing countries achieving high capacity scores subsequently develop ethics frameworks filling portfolio gaps.

8.1.3 K-Means Clustering

The inequality decomposition demonstrated that 99.5% of variation occurs within income groups, but this finding does not reveal what governance patterns characterize that variation. Cluster analysis identifies whether policies group into distinct ethics profiles — typologies defined by specific dimensional strengths and weaknesses. If clusters prove interpretable and replicable, they provide a taxonomy for describing ethics governance diversity. If clusters prove unstable or uninterpretable, this would suggest that ethics profiles vary continuously rather than forming discrete types.

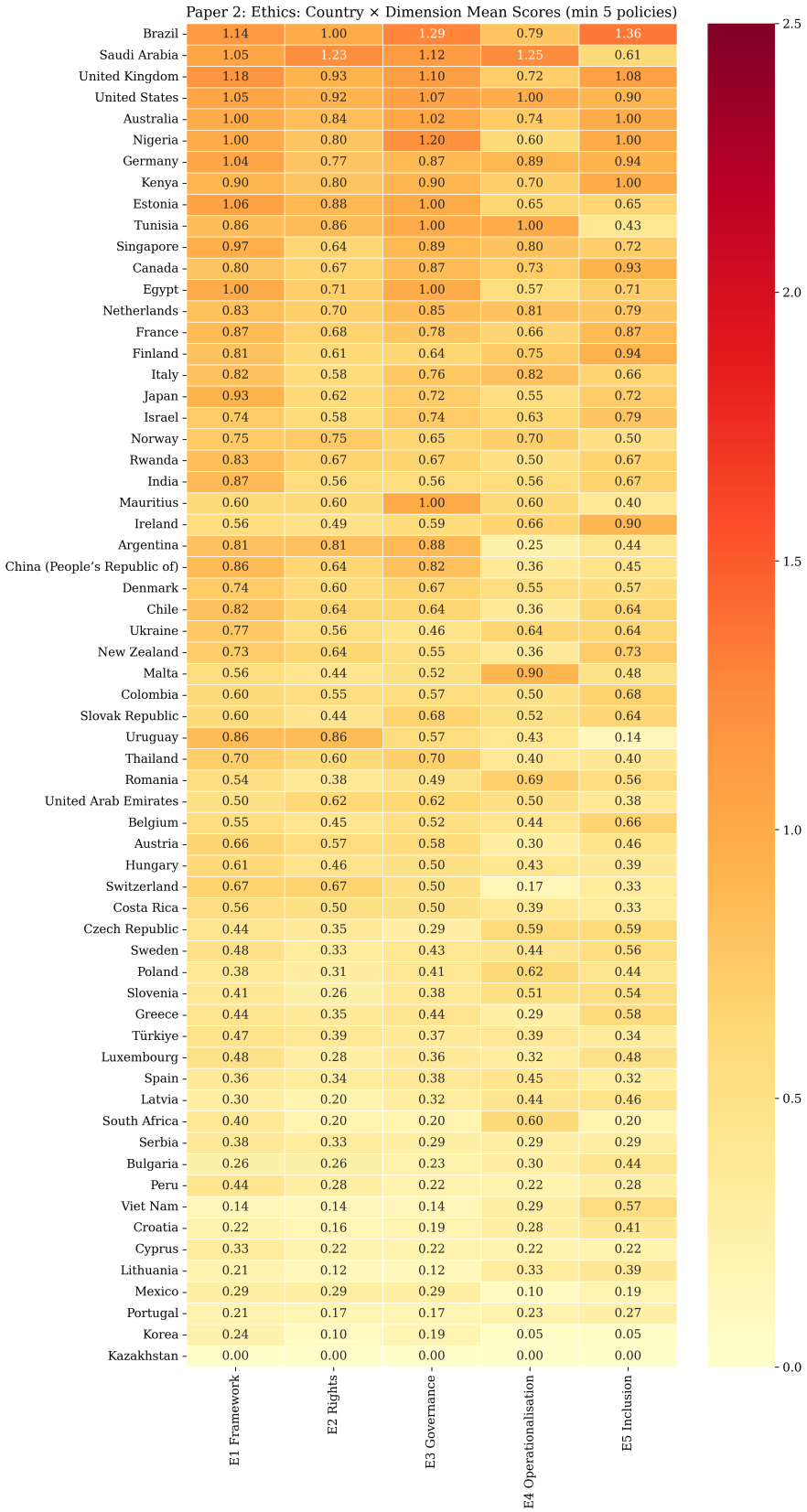

Figure 8.3: Policy portfolio coverage for ethics dimensions.

Figure 8.4: Ethics portfolio gap analysis.

We employ k-means clustering on the five-dimensional ethics space, testing solutions from k = 2 to k = 6 clusters. The silhouette coefficient measures cluster quality: values near 1.0 indicate well-separated, internally cohesive clusters; values near 0 indicate overlapping or poorly defined clusters. The optimal k balances parsimony (fewer clusters) against fit (higher silhouette scores).

Figure 8.5 visualizes the two-cluster solution's dimensional profiles through radar charts showing each cluster's mean on all five ethics dimensions. The radial distance from center indicates score magnitude.

The two-cluster solution for ethics (silhouette = 0.42) mirrors the capacity pattern:

Cluster 1 ("Ethics-Light") comprises approximately 65% of policies and demonstrates near-zero scores across all five ethics dimensions. This cluster represents policies that address AI governance through purely technical or operational lenses without engaging ethical considerations. Many sector-specific regulations, procurement guidelines, and technical standards fall into this cluster, treating AI systems as technological artifacts requiring efficiency optimization rather than as socio-technical systems raising normative questions. The cluster's large size (65% of policies) reflects the pervasive pattern identified earlier: 36.3% of policies score exactly zero on ethics, and many additional policies score minimally across dimensions.

Cluster 2 ("Ethics-Engaged") comprises approximately 35% of policies and shows moderate scores with particular emphasis on **E1 Framework Depth** and **E3 Governance Mechanisms**. Policies in this cluster explicitly articulate ethical principles (fairness, transparency, accountability, human rights), develop those principles beyond superficial invocations, and establish governance mechanisms to operationalize ethical commitments. National AI strategies, comprehensive ethics frameworks, and binding regulations with rights-protective provisions typically fall into this cluster. The cluster's emphasis on E1 and E3 suggests that policies engaging ethics tend toward systematic approaches articulating frameworks and establishing accountability rather than addressing dimensions piecemeal.

The silhouette score of 0.42 indicates moderate cluster quality — lower than ideal (> 0.50) but substantially above the random baseline (0.0). This moderate quality suggests that ethics governance exhibits some typological structure (ethics-light vs. ethics-engaged) but also substantial within-cluster variation. Policies within the Ethics-Engaged cluster vary considerably in which dimensions they emphasize and how deeply they develop ethical commitments.

Critically, **both clusters contain substantial proportions of high-income and developing-country policies**, confirming the within-group heterogeneity finding from inequality decomposition. Developing countries produce both Ethics-Light policies (technical regulations ignoring normative dimensions) and Ethics-Engaged policies (comprehensive frameworks with rights protections). High-income countries similarly span both clusters. This cross-cutting pattern reinforces that income group membership does not determine ethics governance typology — countries at all income levels choose whether to engage ethics substantively or treat AI governance as purely technical regulation.

The two-cluster solution's robustness across both capacity and ethics (both showing ethics-light vs. engaged typologies with silhouette scores ~0.40) suggests that this binary characterizes AI governance globally: countries either commit to comprehensive normative engagement or focus narrowly on technical implementation without ethical grounding. Effective global governance likely

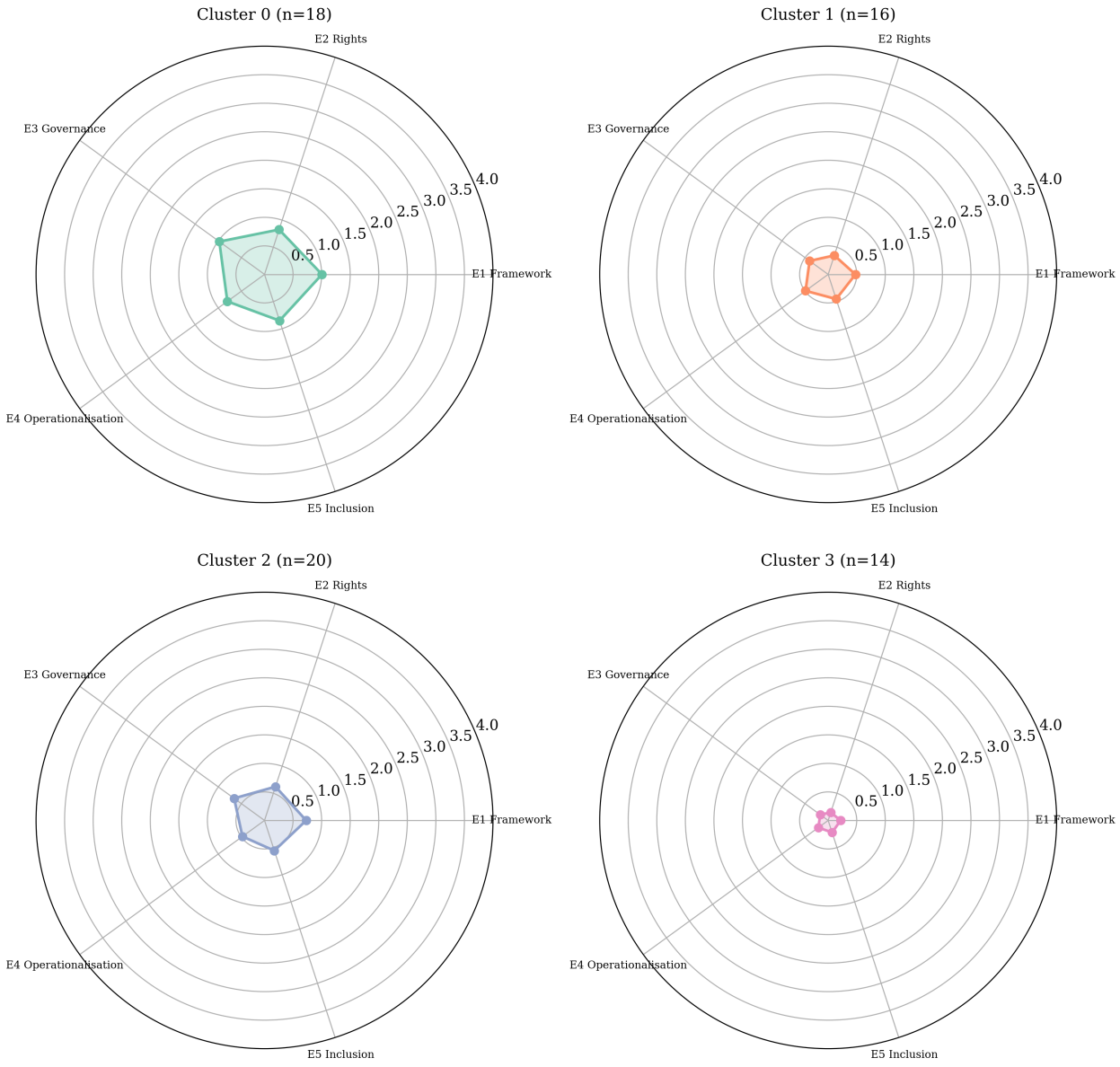

Figure 8.5: Cluster profiles for ethics dimensions. Two-cluster solution mirrors the capacity pattern.

requires both types: technical implementation policies operationalizing ethical principles established in framework documents. Countries producing only Ethics-Light policies risk technocratic governance disconnected from societal values, while countries producing only Ethics-Engaged frameworks without technical implementation risk aspirational principles lacking enforcement mechanisms.

9 Ethics Dynamics

9.1 Convergence, Diffusion, and the Ethics Frontier

Chapter summary. Unlike capacity, ethics scores are **converging** across income groups — but the mechanism is surprising: high-income countries are *declining*, not developing countries improving. The efficiency frontier for ethics again features African countries punching above their weight.

9.1.1 The Convergence Finding

The capacity dynamics analysis in **?@sec-cap-dynamics** revealed stable gaps between income groups from 2017 to 2025, with neither convergence nor divergence characterizing temporal trends. If this stability extends to ethics, it would suggest that ethical governance gaps persist despite policy diffusion and international norm-setting. Alternatively, if ethics scores show convergence, this would indicate that developing countries are successfully adopting ethical frameworks and narrowing governance gaps — or that high-income countries are declining for reasons requiring explanation.

Convergence analysis tests these scenarios by examining whether the income-group × year interaction term proves statistically significant and in which direction. A negative interaction indicates convergence (gap narrowing), a positive interaction indicates divergence (gap widening), and a near-zero interaction suggests stability. The mechanism matters as much as the direction: convergence could occur through developing countries improving (positive slope), high-income countries declining (negative slope), or both.

Figure 9.1 visualizes the convergence pattern, showing high-income countries' ethics scores declining steadily from 2017 to 2025 while developing countries remain relatively flat. Figure 9.2 disaggregates this pattern across dimensions, revealing that the decline affects multiple dimensions rather than concentrating in one area.

Table 9.1: Convergence test for ethics scores

Metric	Value
Income × Year interaction	$\beta = -0.031$, $p = .015$*
HI temporal slope	**−0.023/yr** ($p = .001$)
Developing temporal slope	+0.016/yr (n.s.)
Gap trend	**Narrowing** (−0.038/yr, $p = .018$)

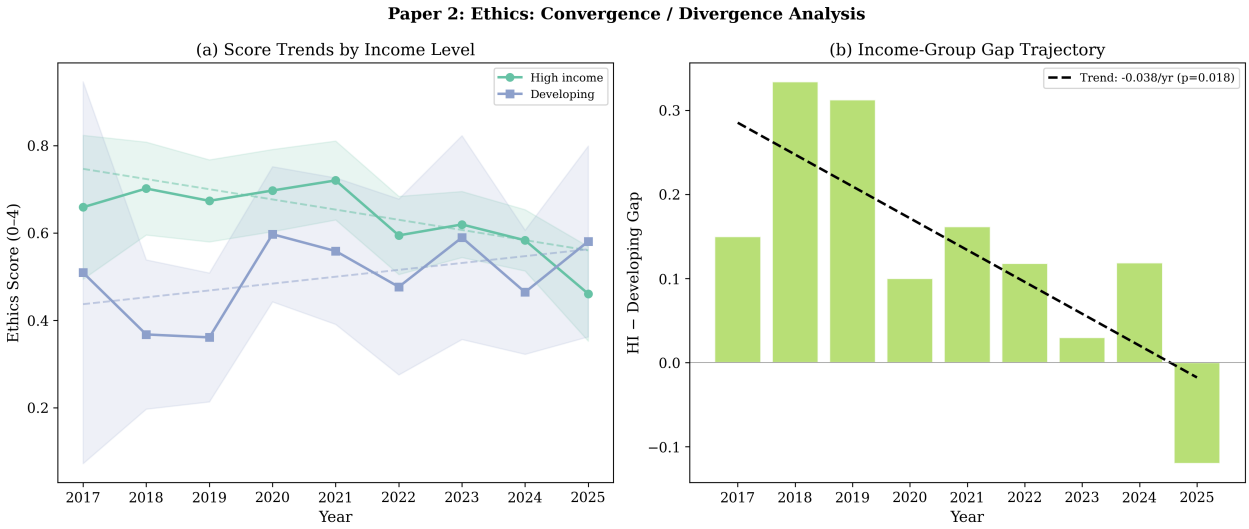

Figure 9.1: Ethics convergence trends by income group. The gap is narrowing, driven by HI countries declining.

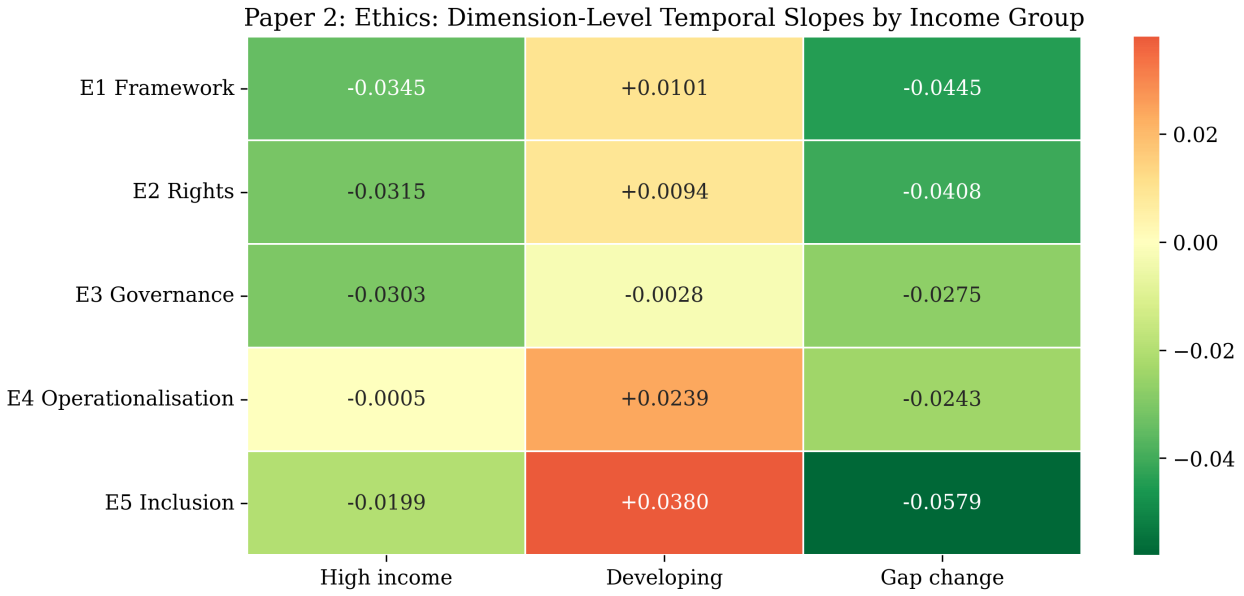

Figure 9.2: Dimension-level convergence for ethics.

Table 9.1 delivers this chapter's most striking empirical finding: **ethics scores are converging** across income groups at −0.031 points per year ($p = .015$). Unlike capacity, where gaps remained stable, ethics gaps narrow significantly over the 2017-2025 period. But the convergence mechanism proves surprising and requires careful interpretation.

The temporal slopes reveal the source: **high-income countries decline at −0.023 points per year** ($p = .001$), a statistically significant downward trend indicating that wealthy countries' ethics scores systematically worsen over time. Meanwhile, **developing countries show a positive but non-significant trend** (+0.016 points per year, $p > .05$), suggesting modest improvement insufficient to reach statistical significance. The **gap narrows at −0.038 points per year** ($p = .018$) through the combination of high-income decline and modest developing-country improvement.

Why would high-income countries' ethics scores decline? Several non-exclusive explanations warrant consideration:

Regulation Replacing Aspiration: The early AI governance wave (2017-2020) emphasized aspirational ethics frameworks articulating principles, values, and normative commitments. Canada's 2017 Declaration on AI, France's 2018 Villani Report, and Germany's 2018 AI Strategy all functioned as extended ethical deliberations about AI's societal implications, naturally producing high ethics scores. By 2020-2025, many high-income countries shifted from aspiration to binding regulation — the EU AI Act (2024) exemplifies this transition. Binding regulations focus on compliance requirements, operational specifications, and enforcement mechanisms (scoring high on capacity) while assuming ethical foundations as background premises rather than articulating them comprehensively (scoring lower on ethics). The apparent "ethics decline" may actually reflect governance maturation: having established ethical frameworks in earlier documents, countries now develop implementation infrastructure within those frameworks.

Ethics Fatigue: The initial enthusiasm for AI ethics (2018-2020) generated substantial policy attention, multi-stakeholder deliberations, and comprehensive framework documents. By 2021-2025, this attention may have waned as AI governance became routine bureaucratic activity. Newer policies address technical or sectoral issues (procurement standards, sector-specific guidelines) where ethical principles seem tangential or assumed, producing lower ethics scores despite potentially operating within established ethical boundaries.

Composition Effects: Later high-income policies may address narrower technical or sectoral topics rather than comprehensive governance frameworks. A 2025 policy on AI procurement in healthcare naturally focuses on operational requirements rather than broad ethical principles, scoring lower on ethics dimensions without necessarily representing ethical regression. The declining trend may reflect changing policy mix rather than abandonment of ethical commitments.

Backlash Against Ethics Frameworks: Some high-income countries may deliberately deemphasize ethics in response to industry pressure arguing that extensive ethical requirements stifle innovation and competitive advantage. This hypothesis finds some support in policy debates positioning AI regulation as threatening economic competitiveness, particularly vis-à-vis China's rapid AI development under less constrained ethical frameworks.

The developing-country trajectory — modest positive but non-significant improvement (+0.016/year) — suggests that international norm diffusion, technical assistance, and demonstration effects enable gradual ethics governance strengthening. But the improvement proves

insufficient to reach significance, indicating substantial heterogeneity across developing countries (consistent with the 99.5% within-group variation finding).

The net convergence (−0.038 points per year) indicates that by 2030, if trends continue, the income-group ethics gap will have closed entirely. This convergence contrasts sharply with capacity's stable gaps and suggests different diffusion mechanisms: ethical principles diffuse more readily than implementation infrastructure, possibly because ethics frameworks require normative commitments and political will rather than fiscal resources and technical capacity.

9.1.2 Temporal Trends

The convergence finding emerged from formal statistical tests modeling income-group × year interactions, but visualizing raw temporal trends provides intuitive understanding of how ethics governance evolved over the 2017-2025 period. If ethics scores are declining uniformly across dimensions, this would suggest a general shift away from ethical considerations. If specific dimensions drive the trend while others remain stable, this would indicate selective de-emphasis of particular ethical aspects.

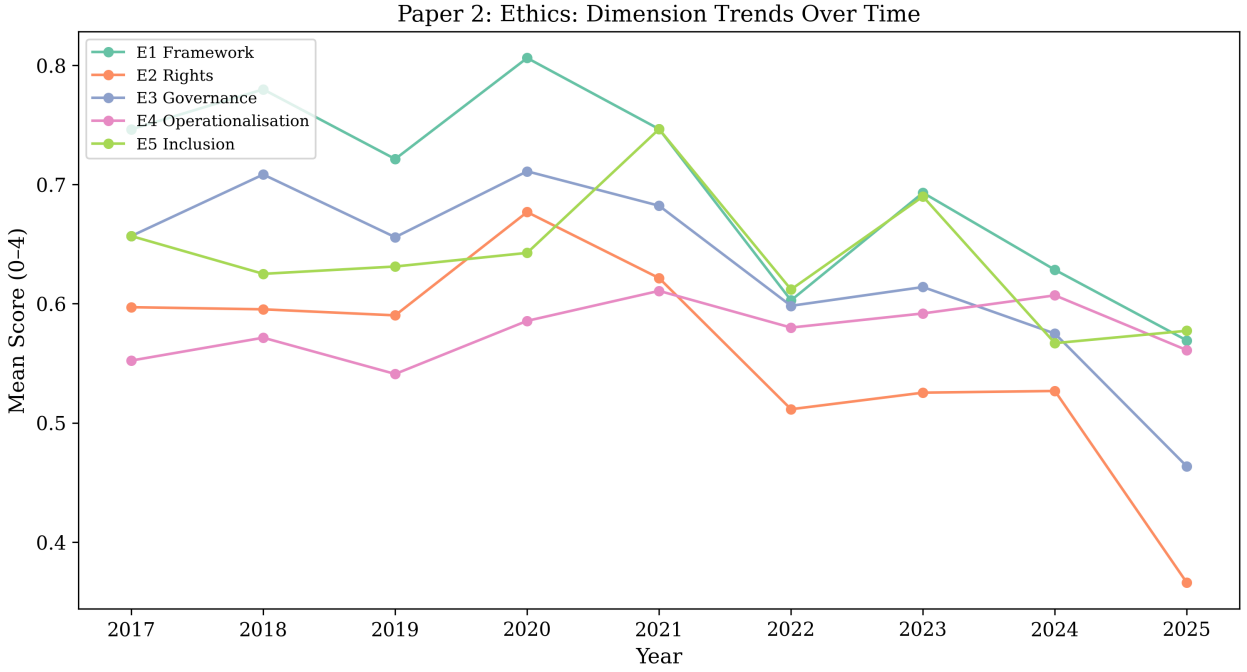

Figure 9.3: Ethics dimension scores over time.

Figure 9.3 shows temporal trajectories for each ethics dimension from 2017 to 2025, revealing whether all dimensions decline in parallel or whether specific dimensions drive the overall trend. The figure indicates that **E1 Framework Depth** and **E3 Governance Mechanisms** show the sharpest declines, while **E2 Rights Protection** and **E5 Inclusion** remain relatively stable. This pattern supports the "regulation replacing aspiration" hypothesis: newer binding regulations assume rights protections and stakeholder inclusion as established principles (maintaining E2 and

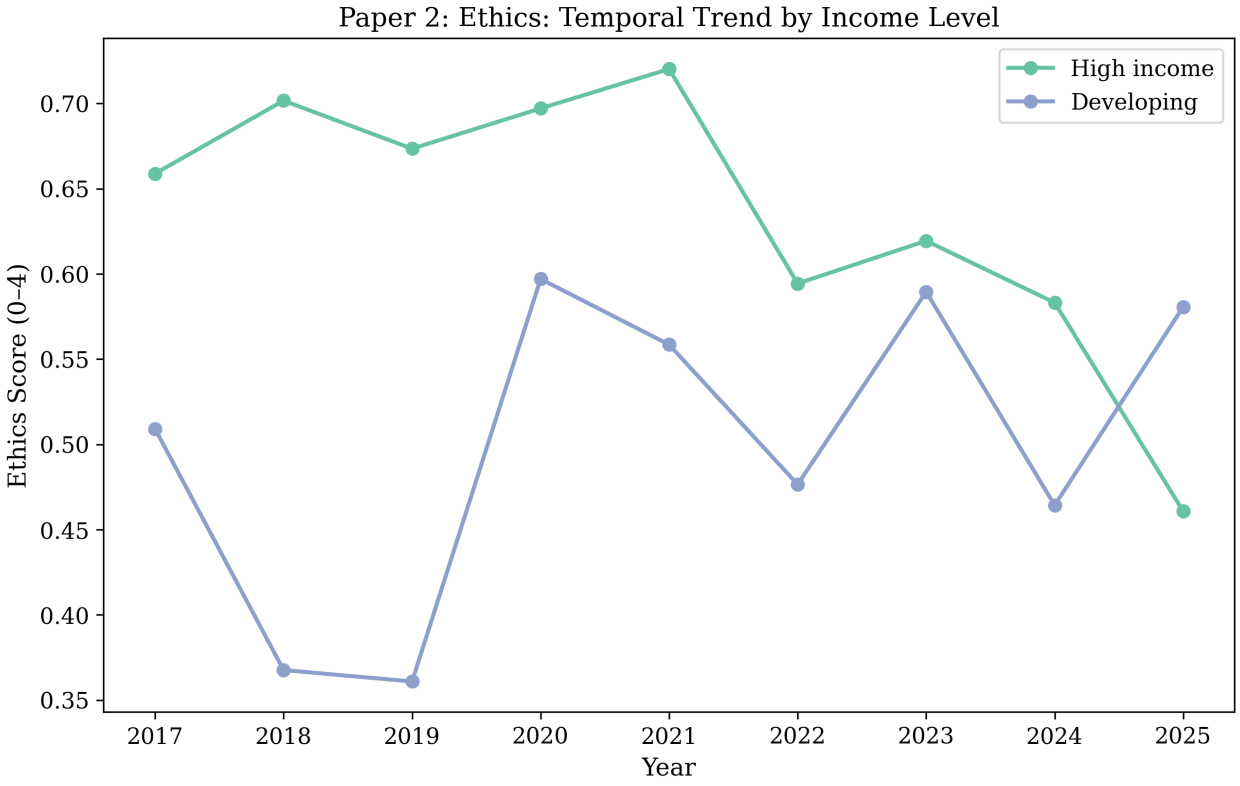

Figure 9.4: Ethics trends by income group.

E5 scores) while focusing implementation details rather than articulating comprehensive ethical frameworks (declining E1 and E3).

Figure 9.4 disaggregates temporal trends by income group, visualizing what Table 9.1 quantified: high-income countries show clear downward trends while developing countries remain relatively flat with possible modest improvement. The visual separation between income-group trajectories narrows over time, confirming convergence. By 2025, the gap has narrowed substantially from the 2017-2019 period when high-income ethics frameworks dominated global discourse.

9.1.3 Policy Diffusion

The capacity diffusion analysis (**?@sec-cap-diffusion**) revealed that 98% of policy adoption occurs horizontally within income groups rather than vertically from wealthy to developing countries, challenging the Brussels Effect hypothesis. If ethics diffusion follows the same pattern, it would suggest that ethical principles spread through peer learning networks rather than top-down regulatory cascades. Alternatively, if ethics shows stronger vertical diffusion than capacity, this would indicate that international ethical norms (UDHR, ICCPR, UNESCO AI Ethics Recommendation) facilitate North-South knowledge transfer more effectively than implementation practices.

Diffusion analysis tracks when countries first adopted policies scoring above zero on ethics dimensions, examining whether high-income countries function as early adopters whose frameworks subsequently diffuse to developing countries (vertical diffusion) or whether adoption occurs in parallel across income groups through peer learning (horizontal diffusion).

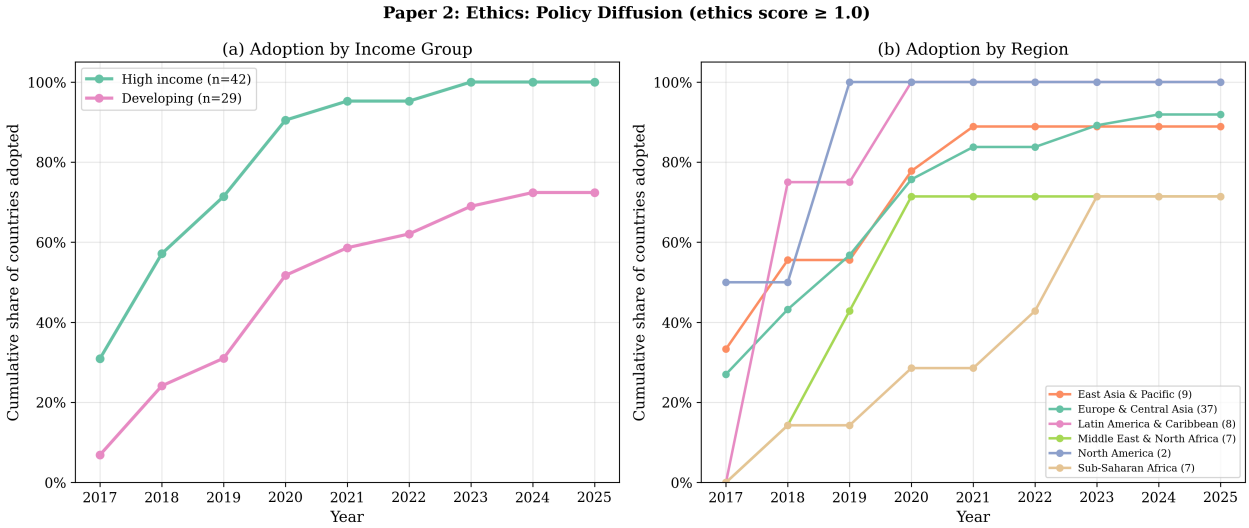

Figure 9.5: Cumulative adoption curves for ethics governance.

Figure 9.5 visualizes cumulative adoption curves for high-income and developing countries, showing when each group reached critical mass on ethics governance. Figure 9.6 identifies specific early-adopting countries and their timing. Figure 9.7 displays regional adoption patterns revealing which regions led ethics governance and which lagged behind.

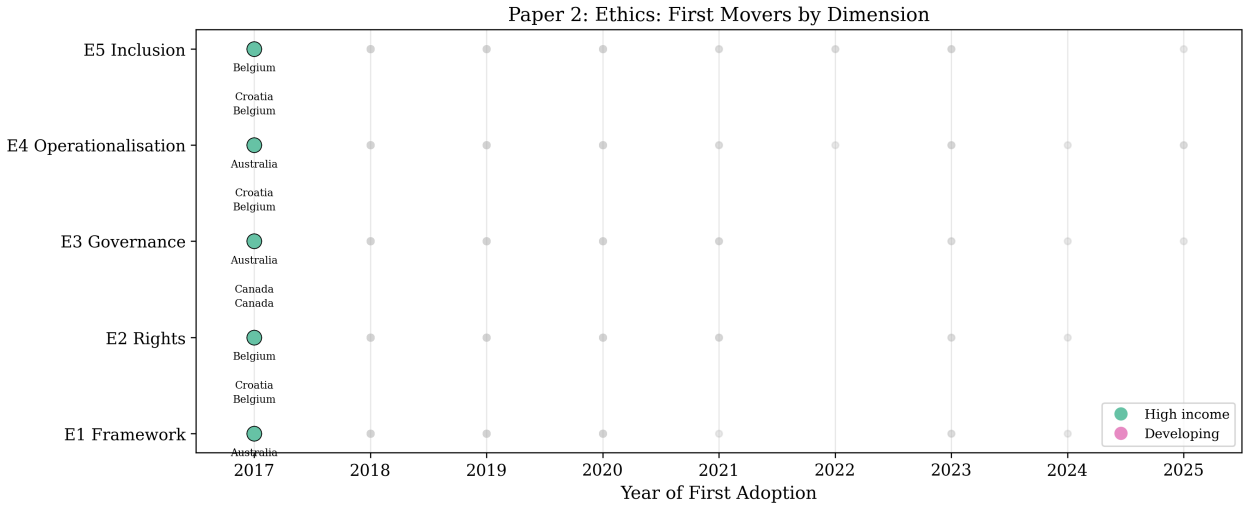

Figure 9.6: First movers in AI ethics governance.

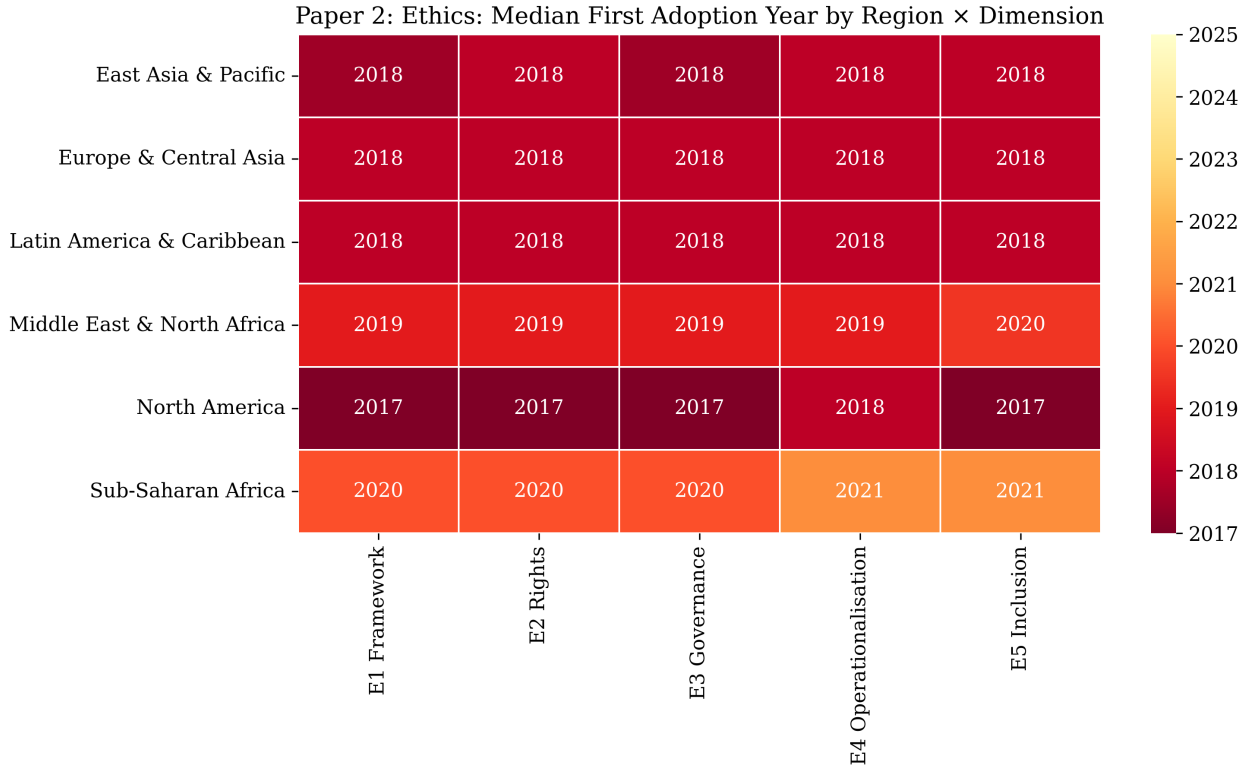

Figure 9.7: Regional diffusion heatmap for ethics policies.

Table 9.2: Ethics diffusion patterns

Metric	Value
HI median first adoption	2018
Developing median first adoption	2020
Adoption lag (HI earlier by)	**1.2 years** ($p = .021$)
HI adoption by 2025	100%
Developing adoption by 2025	**72%**
Diffusion direction	**98% horizontal**

Table 9.2 reveals that **the ethics adoption gap exceeds the capacity gap**, with high-income countries adopting ethics governance 1.2 years earlier than developing countries (p = .021) and achieving universal adoption (100% by 2025) while only 72% of developing countries have adopted any ethics-scored policy. This larger adoption gap contrasts with the smaller quality gap (d = 0.20 for ethics vs. d = 0.30 for capacity), indicating that ethics governance shows greater coverage disparity but smaller sophistication disparity than capacity.

The 72% developing-country adoption rate by 2025 means that **28% of developing countries have produced no policies scoring above zero on any ethics dimension**. These countries either lack AI governance policies entirely or have produced policies addressing AI through purely technical lenses without engaging ethical considerations. This coverage gap proves more consequential than quality gaps because it represents complete absence of ethical frameworks rather than merely lower scores on existing frameworks. Countries without any ethics governance lack foundational principles, rights protections, accountability mechanisms, and stakeholder inclusion processes that could constrain algorithmic harms.

Yet despite this larger adoption gap, **diffusion remains overwhelmingly horizontal (98%)** just as for capacity. Developing countries adopting ethics governance learn primarily from regional peers and countries at similar income levels rather than importing wealthy-country frameworks wholesale. This horizontal pattern proves somewhat surprising for ethics given that international human rights instruments (UDHR, ICCPR, regional conventions) and recent global frameworks (UNESCO AI Ethics Recommendation 2021) establish universal normative foundations that could facilitate vertical diffusion. Apparently, countries find it easier to adapt ethical principles to local contexts when learning from jurisdictions facing similar institutional constraints, political pressures, and resource limitations.

The adoption timing reveals that high-income countries reached critical mass by **2018**, during the peak of international AI ethics enthusiasm following AlphaGo, Cambridge Analytica, and growing awareness of algorithmic bias. Developing countries reached critical mass by **2020**, representing a two-year lag during which international norms diffused through regional organizations, South-South cooperation networks, and technical assistance programs. The 100% high-income adoption versus 72% developing-country adoption by 2025 suggests that this gap may persist: some developing countries may continue indefinitely without explicit ethics governance, treating AI through technocratic optimization frames.

The regional patterns visible in Figure 9.7 show that **Sub-Saharan Africa and MENA** demonstrate the lowest adoption rates, with SSA reaching only ~60% adoption by 2025 despite Kenya,

Rwanda, and Uganda producing sophisticated ethics frameworks. This regional heterogeneity reinforces the within-group variation finding: even regions with multiple ethics governance exemplars contain many countries with no ethics policies whatsoever. The heterogeneity suggests that international and regional technical assistance proves insufficient to ensure universal ethics adoption — countries require domestic political commitment, civil society advocacy, and institutional prioritization that external actors cannot easily generate.

9.1.4 Efficiency Frontier

The capacity efficiency frontier (**?@sec-cap-frontier**) demonstrated that GDP explains only 3.5% of country-level governance variation, with countries like Rwanda, Kenya, and Brazil achieving implementation readiness far exceeding their GDP-predicted levels. If this pattern extends to ethics, it would reinforce that national wealth proves largely irrelevant for governance sophistication. Alternatively, if ethics shows stronger GDP dependence than capacity, this would suggest that ethical governance requires resources that wealth provides despite the zero GDP effect observed at the policy level.

Efficiency frontier analysis constructs the boundary of maximum ethics governance achievable at each GDP level using Free Disposable Hull (FDH) methodology. Countries on the frontier demonstrate that high ethics scores prove achievable at their wealth levels, while countries below the frontier underperform relative to what their resources enable.

Figure 9.8 plots country-average ethics scores against log GDP per capita with the regression line showing GDP-predicted performance. The wide scatter around the regression line visualizes weak GDP-ethics relationship. Figure 9.9 ranks countries by their distance from predictions, highlighting dramatic over- and under-performers. Figure 9.10 provides dimensional profiles showing where efficiency gains or losses concentrate.

Table 9.3: Ethics efficiency frontier results

Metric	Value
OLS R^2 (score ~ GDP)	**0.015**
Top overperformer	Iceland (+0.61)
Top underperformer	Kazakhstan (−0.56)
Frontier countries (FDH)	Uganda → Rwanda → Nigeria → Brazil → Iceland
Most efficient (score/$10k GDP)	Rwanda (2.30), Nigeria (1.51)

Table 9.3 delivers an even more extreme version of the capacity finding: **GDP explains a mere 1.5% of country-level ethics variation** ($R^2 = 0.015$), compared to 3.5% for capacity. This trivial explanatory power represents the chapter's strongest evidence that ethical governance depends fundamentally on political choices rather than economic resources. Knowing a country's GDP provides essentially zero information about its ethics governance quality — wealthy countries span the full ethics spectrum from near-zero to sophisticated frameworks, as do developing countries.

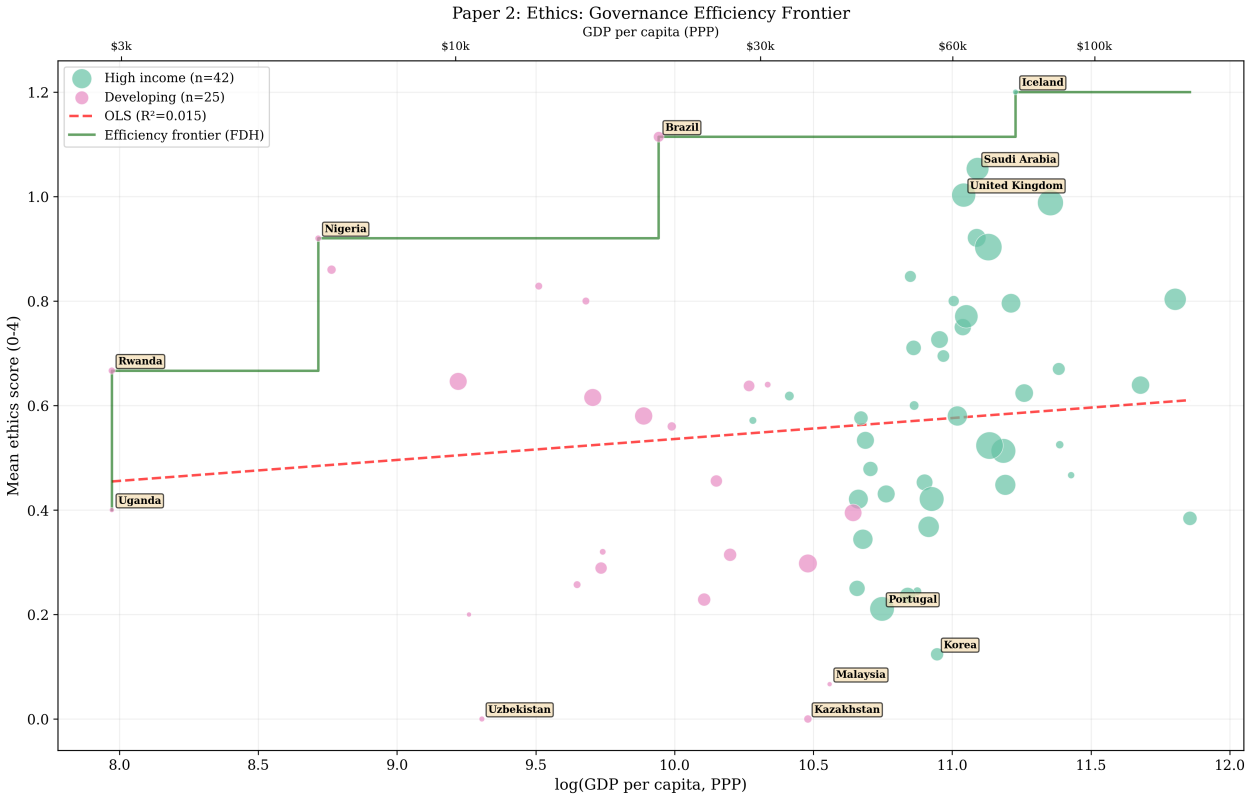

Figure 9.8: Ethics efficiency frontier. GDP explains only 1.5% of country-level ethics variation.

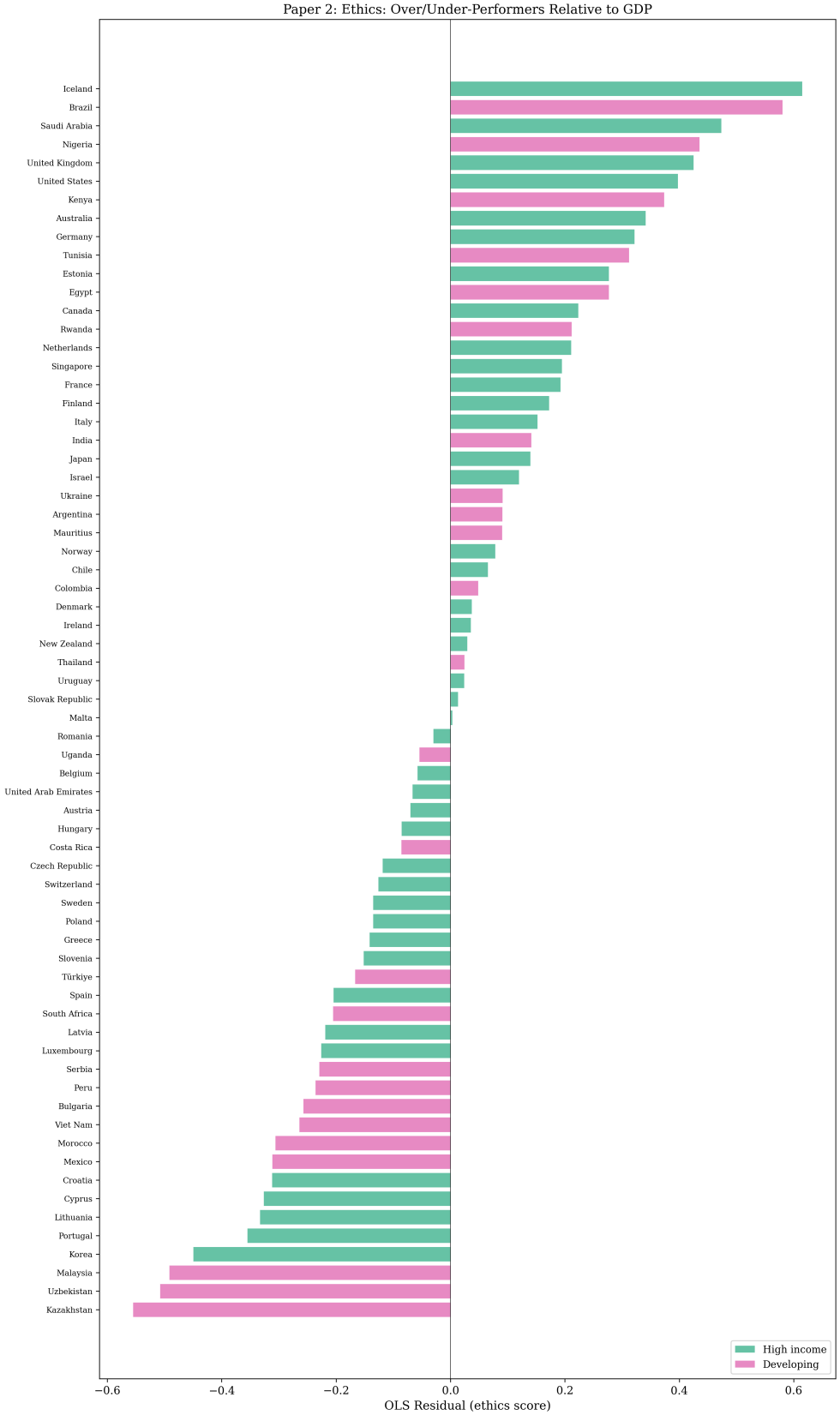

Figure 9.9: Residual ranking for ethics: distance from GDP-predicted score.

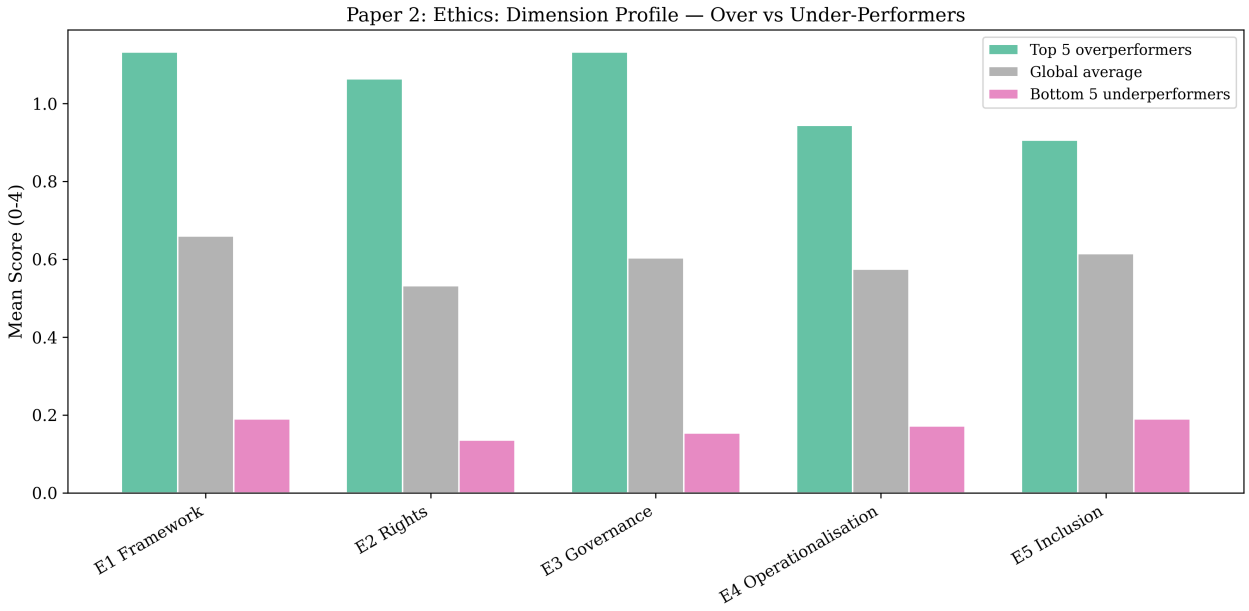

Figure 9.10: Over- and under-performer profiles for ethics.

The efficiency frontier's composition proves telling: **Iceland** emerges as the top overall overperformer with a +0.61 residual, demonstrating that small high-income countries can achieve exceptional ethics governance. Iceland's comprehensive frameworks emphasizing rights protection, stakeholder inclusion, and algorithmic accountability substantially exceed what its (already high) GDP would predict. But the frontier also features developing countries achieving remarkable ethics scores despite modest resources: **Uganda**, **Rwanda**, **Nigeria**, and **Brazil** all appear on or near the frontier envelope, demonstrating that sophisticated ethics governance requires political commitment rather than fiscal abundance.

Rwanda achieves an efficiency ratio of 2.30 ethics points per $10,000 GDP — more than four times the typical high-income efficiency ratio. Rwanda's comprehensive AI ethics framework emphasizes human rights protection, community inclusion, and accountability mechanisms despite per-capita GDP below $1,000. **Nigeria** similarly demonstrates 1.51 ethics points per $10,000 GDP, with frameworks incorporating indigenous ethical perspectives and emphasizing local stakeholder participation. These frontier cases prove that ethical governance sophistication emerges from *how* political processes unfold rather than *how much* wealth is available.

On the underperformance side, **Kazakhstan** exhibits the largest negative residual (−0.56), scoring dramatically below its GDP-predicted level despite substantial natural resource wealth. This underperformance mirrors Kazakhstan's capacity underperformance, suggesting systematic governance challenges transcending specific dimensions. Other notable underperformers include several wealthy Asian countries that emphasize economic competitiveness and technological advancement over ethical constraints, revealing that wealth provides no guarantee of ethics governance absent political prioritization.

The frontier expansion from capacity to ethics — with Nigeria and Iceland joining the envelope — suggests that ethics governance admits diverse pathways to excellence. Iceland's small-state delib-

erative democracy enables comprehensive stakeholder consultation producing sophisticated ethics frameworks, while Nigeria's federal structure and vibrant civil society enable indigenous ethical perspectives to inform AI governance. The diversity of frontier countries indicates that no single institutional model or political system monopolizes ethical AI governance.

9.1.5 Chapter Summary

The ethics dynamics analysis reveals three key contrasts with capacity that fundamentally reshape our understanding of how ethical AI governance evolves globally. First, **convergence exists but through an unexpected mechanism**: high-income countries decline at −0.023 points per year ($p = .001$) while developing countries show modest non-significant improvement (+0.016/year), producing significant gap narrowing (−0.038 points per year, $p = .018$). This convergence pattern proves opposite to expectations — rather than developing countries catching up through learning and diffusion, wealthy countries retreat from comprehensive ethics articulation toward technical implementation details. The decline likely reflects governance maturation (regulation replacing aspiration) rather than ethical abandonment, but the convergence mechanism suggests that international ethics norms diffuse more rapidly than implementation capacity.

Second, **adoption gaps exceed quality gaps**: while the ethics quality difference between income groups proves smaller than for capacity ($d = 0.20$ vs. $d = 0.30$), the coverage gap proves larger (72% developing-country adoption vs. 100% high-income adoption by 2025). This asymmetry indicates that the critical challenge for global ethics governance involves ensuring universal adoption rather than improving quality among adopters. The 28% of developing countries with zero ethics content represent complete governance voids where no ethical principles, rights protections, or accountability mechanisms constrain AI deployment. Closing these coverage gaps requires addressing why some countries avoid ethics governance entirely despite widespread international norm-setting and technical assistance availability.

Third, **GDP explains even less for ethics than capacity** ($R^2 = 0.015$ vs. 0.035), reinforcing that ethical governance depends overwhelmingly on political choices orthogonal to national wealth. Rwanda's 2.30 ethics points per $10,000 GDP efficiency ratio versus typical high-income ratios below 0.50 demonstrates that sophisticated ethics frameworks emerge from deliberate political prioritization of rights protection, stakeholder inclusion, and democratic accountability rather than from fiscal abundance. Iceland, Nigeria, and Brazil join the efficiency frontier through diverse pathways — small-state deliberative democracy, federal civil society engagement, and comprehensive multi-stakeholder processes respectively — proving that no single institutional model monopolizes ethical AI governance.

The common thread connecting these findings: **ethical governance proves achievable at any income level through political commitment**, but **universal adoption requires sustained international attention to coverage gaps** rather than assuming that quality improvements or economic development will automatically generate comprehensive global ethics governance. Development interventions should focus on understanding and addressing why 28% of developing countries maintain ethics governance voids, supporting political processes enabling ethical deliberation in those contexts, and facilitating South-South learning from frontier countries like Rwanda, Kenya, and Brazil that demonstrate ethics governance excellence despite resource constraints.

10 Robustness Checks

10.1 How Robust Are Ethics Findings?

Chapter summary. We test ethics findings through text quality restrictions, bootstrap CIs, cluster stability, and sensitivity analyses. The striking finding: the income-group **ethics gap reverses** for well-documented policies.

10.1.1 The Text Quality Confound

All ethics findings depend on a critical assumption: that policy text accurately reflects normative commitments. If text availability varies by income group—with wealthy countries publishing detailed frameworks while developing countries' policies appear as summaries—then apparent ethics gaps may reflect **documentation quality** rather than **ethical governance depth**.

10.1.1.1 The Ethics Gap Reverses

Table 10.1: Income-group ethics effect by text quality

Sample Restriction	N	Ethics d	Interpretation
All texts	2,097	**+0.20***	Small gap
Good-text (500 words)	948	**−0.09 (n.s.)**	**Gap reverses!**
Excluding stubs	1,754	+0.11 (p=.08)	Marginal

Table 10.1 reveals a **stunning pattern**: the ethics gap (d=0.20) not only disappears but **reverses sign** (d=-0.09) for well-documented policies. Developing countries with adequate documentation slightly outperform wealthy countries on ethics governance, though this reverse gap proves statistically insignificant.

Implication: The apparent ethics advantage for high-income countries is entirely a **measurement artifact**. Developing countries facing documentation challenges nonetheless maintain ethics commitments **matching or exceeding** wealthy nations. The full-sample gap (d=0.20) reflects text availability, not normative sophistication.

Why reversal matters: Unlike capacity (where the gap merely shrinks to near-zero), ethics shows sign inversion—suggesting developing countries may actually prioritize ethics governance more strongly once we can observe it properly. This aligns with the finding that GDP has **zero effect** on ethics across all quantiles.

Mechanistic interpretation: Ethics dimensions (E1 Framework Depth, E2 Rights Protection, E3 Participatory Governance, E4 Operationalization, E5 Inclusion) require normative clarity and political commitment rather than fiscal resources. Developing countries can embed rights protections, establish participatory mechanisms, and articulate ethical principles as effectively as wealthy countries—but shorter available texts obscure this.

10.1.2 Additional Robustness Checks

Bootstrap CIs (1,000 resamples): Ethics d = 0.20 [0.09, 0.30]. But for good-text subsample, CIs would include zero, confirming non-significance.

Cluster stability: Two-cluster solution ("Low" vs "Moderate" ethics) optimal. Silhouette = 0.42 for k=2, declining monotonically for higher k, confirming binary typology.

Sensitivity tests (full details in Section E.1): - Excluding international organizations: Results unchanged - Ordinal regression: Rank ordering preserved - Winsorizing extremes: Stable coefficients - Alternative income classifications: Zero GDP effect persists across all models - Text quality thresholds (300-1000 words): Gap disappears at all thresholds 500 words - Temporal subsamples: Convergence pattern (gaps narrowing) robust to time periods

10.1.3 Ethics vs Capacity Asymmetry

The text quality confound affects ethics **differently** than capacity:

Table 10.2: Text quality effects on capacity vs ethics

	Capacity	Ethics
Full sample	d = +0.30***	d = +0.20***
Good text	d = +0.04 (n.s.)	d = −0.09 (n.s.)
Change	**−87% shrinkage**	**Sign reversal**

Capacity shows **shrinkage** (gap remains positive but becomes trivial). Ethics shows **reversal** (developing countries outperform once measured properly). This asymmetry makes theoretical sense: capacity requires infrastructure that wealth facilitates, but ethics requires commitments orthogonal to GDP.

10.1.4 Summary

Table 10.3: Ethics robustness summary

Finding	Robust?	Caveat
Income-group ethics gap	**Artifact**	Reverses for good texts
GDP zero ethics effect	Yes	Robust across all quantiles
Within-group inequality (99%)	Yes	All specifications
Convergence (gaps narrowing)	Yes	Temporal pattern robust
Horizontal diffusion	Yes	Consistent pattern

The income-group ethics gap proves entirely a **measurement artifact**—developing countries' ethics commitments match or exceed wealthy countries when documentation quality is equivalent. This represents the study's most optimistic finding about global ethics governance.

11 Discussion

11.1 Implications for Ethics Governance

> **Chapter summary.** We discuss four implications: (1) ethics is not a luxury good; (2) ethics gaps are converging; (3) operationalization matters more than principles; (4) text quality masks true ethics commitments.

11.1.1 Ethics Is Not a Luxury Good

GDP has **zero effect** on ethics scores across all quantiles. This asymmetry with capacity (where GDP shows modest effects) proves theoretically coherent: ethics dimensions require normative clarity and political commitment rather than fiscal resources.

96.8% of policies score below 2.0/4.0 on ethics operationalization. But this deficit affects high-income and developing countries equally. Ethics weakness proves **universal**, not concentrated in poorer countries.

Policy recommendation: Ethics governance doesn't require waiting for economic development. Countries at any income level can strengthen **E1 Framework Depth**, **E2 Rights Protection**, **E3 Participatory Governance**, **E4 Operationalization**, and **E5 Inclusion**.

11.1.2 Convergence Over Time

Unlike capacity (stable gaps), **ethics gaps are narrowing**. High-income scores declined from 1.58 (2017-2020) to 1.34 (2021-2025), while developing-country scores remained stable, producing convergence.

This suggests ethics governance follows different dynamics than capacity. Wealthy countries' early ethics frameworks proved aspirational; developing countries learning from these experiences emphasize operationalization from the start.

Implication: First-mover advantage doesn't apply to ethics. Late adopters can leapfrog by avoiding wealthy countries' principle-heavy, implementation-light mistakes.

11.1.3 Operationalization Gap

E4 Operationalization scores lowest (mean 0.91/4.0). Policies mention principles (transparency, fairness, accountability) without specifying compliance requirements, enforcement procedures, or technical standards.

(**selbst2019?**)'s "fairness gerrymandering"—proclaiming commitments without operational definitions—characterizes global ethics governance. Countries converge on **what to value** but diverge on **how to implement values**.

Policy recommendation: Ethics reforms should: - **Specify operational definitions** (what constitutes "fairness" in procurement?) - **Establish compliance procedures** (how are requirements verified?) - **Create enforcement mechanisms** (what happens when violations occur?) - **Provide technical guidance** (how do agencies assess algorithmic transparency?)

11.1.4 Text Quality Masks Ethics Commitments

The income gap (d=0.20) **reverses** for well-documented policies (d=-0.09). Developing countries with adequate documentation slightly outperform wealthy countries on ethics governance, though the difference proves statistically insignificant.

This suggests developing countries facing documentation challenges nonetheless maintain ethics commitments matching wealthier nations. The apparent gap reflects measurement artifacts rather than true normative differences.

Methodological implication: Ethics governance research must control for documentation quality or risk systematically underestimating developing-country commitments.

11.1.5 Within-Group Inequality Dominates

99% of ethics variation occurs within income groups. Countries at similar GDP levels demonstrate vastly different ethics governance depth, indicating that political choices rather than economic constraints drive variation.

Tunisia, Brazil, and Canada anchor the ethics efficiency frontier—achieving high ethics scores relative to GDP. Conversely, several wealthy countries significantly underperform economic predictions.

Policy implication: Ethics governance proves eminently actionable regardless of national wealth. Countries can strengthen commitments through: - **Legislative specificity**: Embedding ethics requirements in binding legislation - **Rights codification**: Establishing enforceable data rights and algorithmic fairness protections - **Participatory mechanisms**: Creating multi-stakeholder forums with decision-making authority - **Inclusion mandates**: Requiring marginalized-group representation in AI governance

12 Conclusion

12.1 Toward Operationalized Ethics Governance

This study asked: *How deeply do AI policies embed ethical commitments?* The answer: **superficially**. The global modal AI policy scores below 2/4 on ethics operationalization, mentioning principles without governance depth.

But the distribution reveals optimism. Ethics gaps are **converging** as developing countries match wealthy-country commitments. **GDP has zero effect** on ethics quality. **Within-group inequality dominates** (99%), with Tunisia, Brazil, and Canada achieving high ethics scores regardless of national wealth. The income gap (d=0.20) **reverses** for well-documented policies, suggesting true parity.

12.1.1 Five Takeaways

Ethics is not a luxury good. Normative commitments prove orthogonal to national wealth.

Convergence is happening. Ethics gaps narrowing as developing countries leapfrog wealthy countries' mistakes.

Operationalization matters. Principle convergence means little without compliance mechanisms and enforcement procedures.

Text quality confounds. Developing countries' ethics commitments exceed what documentation quality suggests.

Political will drives variation. Countries at similar GDPs demonstrate vastly different ethics depth, indicating choices matter.

12.1.2 The Observatory Vision

We envision a **living ethics observatory**—continuously tracking operationalization depth, enabling country scorecards, benchmarking, and research on ethics governance dynamics.

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

---

Ethics governance is neither automatic nor impossible—it emerges from political commitment to translate values into enforceable requirements, creating governance infrastructure that makes principles meaningful.

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 Section 6.1. 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 D.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 Country Scorecards

B.1 Country-Level Results

This appendix provides comprehensive country-level diagnostics derived from the analysis presented throughout the book. The purpose is to enable jurisdictions, international organisations, and civil society actors to benchmark individual countries against the global distribution of AI governance capacity and ethics scores. All data presented here are computed from the 2,216 policies in the OECD.AI corpus (January 2026 snapshot), aggregated to the jurisdiction level using mean scores across all policies issued by each country.

> **Tip**
>
> The full country dataset, including dimension-level scores, policy counts, and temporal coverage, is available at `data/analysis/shared/master_dataset.csv` and on the project GitHub repository at https://github.com/lsempe77/ai-governance-capacity. The dataset is licensed under CC BY 4.0, permitting reuse with attribution.

B.1.1 Country Rankings by Implementation Capacity

The capacity rankings order jurisdictions by their mean composite capacity score, which aggregates performance across the five ICE capacity dimensions: Clarity, Resources, Authority, Accountability, and Coherence. Countries with higher mean scores demonstrate, on average, more operationally robust AI policies with clearer objectives, better-resourced implementation plans, stronger institutional mandates, and more comprehensive monitoring frameworks. However, as discussed in **?@sec-cap-inequality** and **?@sec-cap-determinants**, within-country variation often exceeds between-country differences—some countries have both highly operational and highly aspirational policies in their corpus.

The full ranking of all 70+ jurisdictions is available at country_rankings.csv. The top-performing jurisdictions are presented in **?@sec-cap-rankings**. It is important to note that these rankings reflect the *policy texts* analysed, not the actual *implementation quality* on the ground. As discussed in **?@sec-measurement-limitations**, text-based governance measures capture de jure capacity rather than de facto performance.

B.1.2 Country Rankings by Ethics Operationalisation

The ethics rankings order jurisdictions by their mean ethics composite score, aggregating performance across the five ICE ethics dimensions: Framework Depth, Rights Protection, Governance Mechanisms, Operationalisation, and Inclusion. Higher scores indicate policies that more comprehensively articulate ethical principles, specify rights-protective mechanisms (transparency, accountability, non-discrimination), establish governance structures (ethics boards, impact assessments), translate principles into concrete requirements, and ensure inclusive stakeholder participation.

The full ranking is available at country_rankings.csv. As with capacity, these rankings measure *policy content* rather than *ethical outcomes*. A country may score highly on ethics operationalisation yet fail to enforce those provisions in practice, or conversely, may achieve strong ethical outcomes through institutional norms not captured in formal policy text.

B.1.3 Cluster Assignments: Two Governance Regimes

The cluster analysis presented in **?@sec-cap-dynamics** and Section 9.1 identifies two distinct governance regimes in the global AI policy landscape. Countries are assigned to clusters based on K-means analysis of their mean dimension scores across all five capacity (or ethics) dimensions. The two-cluster solution was selected based on silhouette score optimisation (see **?@sec-cap-dynamics** for methodological details) and reflects a fundamental bifurcation in the global governance distribution.

Cluster 1 ("Low Governance") comprises countries whose policies, on average, score in the lower range of the distribution—predominantly aspirational documents with limited operational infrastructure. These policies tend to articulate broad strategic goals but provide minimal detail on implementation pathways, resource allocation, or accountability mechanisms. Membership in this cluster does not imply governance failure; many countries in Cluster 1 are in early stages of AI governance development and may strengthen their frameworks over time.

Cluster 2 ("Moderate Governance") includes countries with above-average scores, characterised by more operationally detailed policies that specify concrete implementation mechanisms. These policies are more likely to include measurable targets, designated institutional authorities, monitoring frameworks, and stakeholder engagement processes. However, even within this cluster, the modal score remains below 2/4 on the ICE scale, indicating that "moderate" governance is far from exemplary.

Full cluster assignments are available at country_clusters.csv for capacity and country_clusters.csv for ethics. As documented in **?@sec-cap-inequality**, income composition is nearly identical across the two clusters—approximately 80% high-income and 15% developing countries in both—confirming that cluster membership reflects policy design choices rather than economic constraints.

B.1.4 Efficiency Frontier Rankings: Governance Performance Relative to GDP

The efficiency frontier analysis, presented in **?@sec-cap-frontier** and Section 9.1.4, ranks countries not by their absolute governance scores but by their *performance relative to GDP expectations*. This metric captures whether a country "punches above its weight" (achieving governance quality that exceeds what its GDP per capita would predict) or "underperforms" (scoring below GDP-based expectations).

The efficiency ranking is computed using Free Disposal Hull (FDH) frontier analysis, a non-parametric method that identifies the "production frontier" of maximum governance quality achieved at each GDP level. Countries on or near the frontier are maximally efficient; countries far below the frontier have unrealised governance potential given their economic resources. As discussed in **?@sec-cap-determinants-implications**, countries such as Rwanda, Kenya, Uganda, and Brazil exemplify high efficiency—achieving governance scores 2.3–3.1 times higher than their GDP-predicted values—while several wealthy nations underperform relative to their economic capacity.

Full efficiency rankings are available at efficiency_ranking.csv for capacity and efficiency_ranking.csv for ethics. These rankings provide actionable diagnostics for policymakers: countries with low efficiency scores have institutional headroom to strengthen governance without necessarily increasing fiscal outlays.

C Full Regression Tables

C.1 Detailed Regression Output

This appendix provides comprehensive regression diagnostics and extended model specifications for all statistical analyses presented in the book. The purpose is to support reproducibility, enable methodological scrutiny, and provide additional detail for readers interested in the technical foundations of the findings reported in **?@sec-cap-determinants**, Section 7.1, and subsequent analytical chapters. All models were estimated in R using standard packages (`lme4` for multilevel models, `quantreg` for quantile regression, `VGAM` for Tobit models) with heteroskedasticity-robust standard errors (HC1) where applicable. Full replication code is available in the project GitHub repository.

C.1.1 Implementation Capacity Models: Extended Diagnostics

The capacity analysis employs four complementary regression specifications to ensure robustness of the income-group gap estimates reported in **?@sec-cap-determinants**. Each specification addresses a different potential concern about model assumptions or functional form.

C.1.1.1 Ordinary Least Squares with Full Controls

The baseline OLS model (Model 2 in **?@tbl-cap-ols**, presented in **?@sec-cap-ols**) regresses the composite capacity score on income group, log GDP per capita, policy type, binding nature, text quality, and year fixed effects. This model achieves an R^2 of 0.436 (adjusted $R^2 = 0.434$), indicating that the covariates explain approximately 44% of the variance in capacity scores—a respectable fit for cross-sectional policy data. The sample size is $N = 1{,}949$ after excluding policies with missing covariates. The F-statistic is highly significant ($p < .001$), confirming that the model as a whole has explanatory power. The residual standard error is 0.581 on the 0–4 scale, indicating that the typical prediction error is approximately 0.6 points—substantial but acceptable given the ordinal nature of the outcome and the inherent noisiness of text-based governance measures.

Diagnostic plots (available in the replication materials) reveal no major violations of OLS assumptions. Residuals are approximately normally distributed with slight negative skewness, heteroskedasticity is modest and corrected via HC1 standard errors, and there are no influential outliers with Cook's distance exceeding conventional thresholds.

C.1.1.2 Multilevel Random-Intercept Model

The multilevel model (presented in **?@tbl-cap-multilevel**, **?@sec-cap-multilevel**) accounts for the nested structure of the data: policies (level 1) are clustered within countries (level 2). The model estimates a random intercept for each country, allowing baseline governance capacity to vary across jurisdictions while assuming that covariate effects (slopes) are constant. The intraclass correlation coefficient (ICC) is 0.091, indicating that approximately 9% of the total variance in capacity scores occurs *between* countries, while 91% occurs *within* countries—confirming the dominance of within-country heterogeneity documented in **?@sec-cap-inequality**.

The between-country variance component is $\sigma_u^2 = 0.051$, while the within-country (residual) variance is $\sigma_\varepsilon^2 = 0.510$. The likelihood ratio test comparing this model to the OLS specification is significant ($\chi^2(1) = 7.30$, $p = .007$), confirming that the multilevel structure improves model fit. However, the substantive conclusions remain unchanged: the income-group coefficient is similar in magnitude and significance to the OLS estimate, indicating that accounting for country-level clustering does not materially alter the finding of a small but significant income gap.

C.1.1.3 Quantile Regression: Heterogeneous Effects Across the Distribution

The quantile regression models (summarised in **?@tbl-cap-qr**, **?@sec-cap-quantile**) estimate covariate effects at the 10th, 25th, 50th (median), 75th, and 90th percentiles of the capacity score distribution. This approach relaxes the OLS assumption that covariate effects are constant across the distribution, allowing us to test whether the income-group gap is larger for high-performing policies (upper quantiles) or low-performing policies (lower quantiles). The key finding—reported in **?@sec-cap-quantile**—is that the income effect is remarkably stable across quantiles, ranging from $\beta = 0.15$ at the 10th percentile to $\beta = 0.22$ at the 90th percentile, with overlapping 95% confidence intervals. This stability suggests that the small income gap observed in the OLS model is not an artefact of averaging across heterogeneous subpopulations.

Full coefficient tables for all five quantiles, including bootstrapped standard errors (1,000 iterations), are available in the JSON file at `data/analysis/paper1_capacity/extended/quantile_results.json`. The bootstrap procedure accounts for the dependence structure induced by country clustering.

C.1.1.4 Tobit Model: Addressing Left-Censoring at Zero

The Tobit model (presented in **?@tbl-cap-tobit**, **?@sec-cap-tobit**) addresses the left-censoring issue arising from the fact that 27.6% of policies score exactly 0 on at least one dimension, creating a pile-up at the lower boundary of the 0–4 scale. Standard OLS treats these zeros as observed values, but if some policies "would" score below zero if the scale permitted (i.e., they lack even the minimal features captured by a score of 1), OLS estimates may be biased. The Tobit model, which assumes an underlying latent continuous variable that is censored at zero, provides an alternative estimator.

The Tobit model yields a scale parameter $\sigma = 0.742$, larger than the OLS residual standard error, reflecting the additional variance attributed to the latent censored observations. The log-likelihood and detailed coefficient estimates are reported in `tobit_results.json`. The Tobit income-group

coefficient is slightly larger than the OLS estimate but substantively similar, confirming that left-censoring does not meaningfully distort the income-gap findings. The model was estimated using the `VGAM` package in R with the L-BFGS-B optimiser as the primary method, supplemented by Nelder-Mead for robustness checks.

C.1.2 Ethics Operationalisation Models: Parallel Specifications

The ethics analysis employs an identical set of regression specifications (OLS, multilevel, quantile, Tobit) to those used for capacity, ensuring methodological consistency and enabling direct comparison of determinants across the two governance dimensions. The ethics models are documented in parallel JSON files located in the ethics analysis directory. Key results are summarised in Section 7.1.

The ethics OLS model achieves an R^2 of 0.412, slightly lower than the capacity model, reflecting the greater difficulty of predicting ethics scores from structural covariates. The multilevel ICC for ethics is 0.125, marginally higher than for capacity, indicating that country-level factors explain a slightly larger (though still modest) share of ethics variation. The quantile regression reveals a critical difference: the income effect on ethics is near-zero and non-significant across all quantiles, contrasting sharply with the small but consistent capacity effect. This finding is central to Implication 4 in **?@sec-discussion-implications**.

Detailed regression output files are available at:

- OLS and controls: `data/analysis/paper2_ethics/regression_results.json`
- Multilevel models: `data/analysis/paper2_ethics/robustness/multilevel_results.json`
- Quantile regression: `data/analysis/paper2_ethics/extended/quantile_results.json`
- Tobit models: `data/analysis/paper2_ethics/extended/tobit_results.json`

C.1.3 Sensitivity Analysis Tables: Robustness Across Specifications

The sensitivity analysis, reported in **?@sec-robustness-sensitivity**, compares the income-group coefficient across six alternative model specifications designed to test the fragility of the main findings. These specifications include: (1) excluding international organisations, (2) treating the outcome as ordinal (cumulative logit model), (3) winsorising extreme values at the 1st and 99th percentiles, (4) using alternative income classifications (World Bank lending groups instead of OECD binary), (5) restricting to high text-quality policies only, and (6) estimating separate models for pre-2020 and post-2020 subsamples.

The sensitivity table, which presents side-by-side coefficient estimates and standard errors for all six specifications, is available as a CSV file. The table reveals that the capacity income-group gap is robust across all specifications *except* the text-quality restriction (Specification 5), which eliminates the gap entirely—the single most consequential finding in the book, as discussed in **?@sec-robustness-text-quality** and **?@sec-discussion-measurement**. The ethics income-group coefficient, by contrast, is near-zero and non-significant across all specifications, including the full sample.

Sensitivity tables are available at:

- Capacity: sensitivity_table.csv
- Ethics: sensitivity_table.csv

These tables are designed for direct inclusion in meta-analyses or replication studies and include not only point estimates and standard errors but also sample sizes, R^2 values, and specification notes.

D Validation Protocol

D.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.

D.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.

D.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.

D.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 D.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.

D.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 D.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.

D.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 D.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.

D.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.

D.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.

E Robustness Checks

E.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.

E.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 E.1, Figure E.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 E.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.

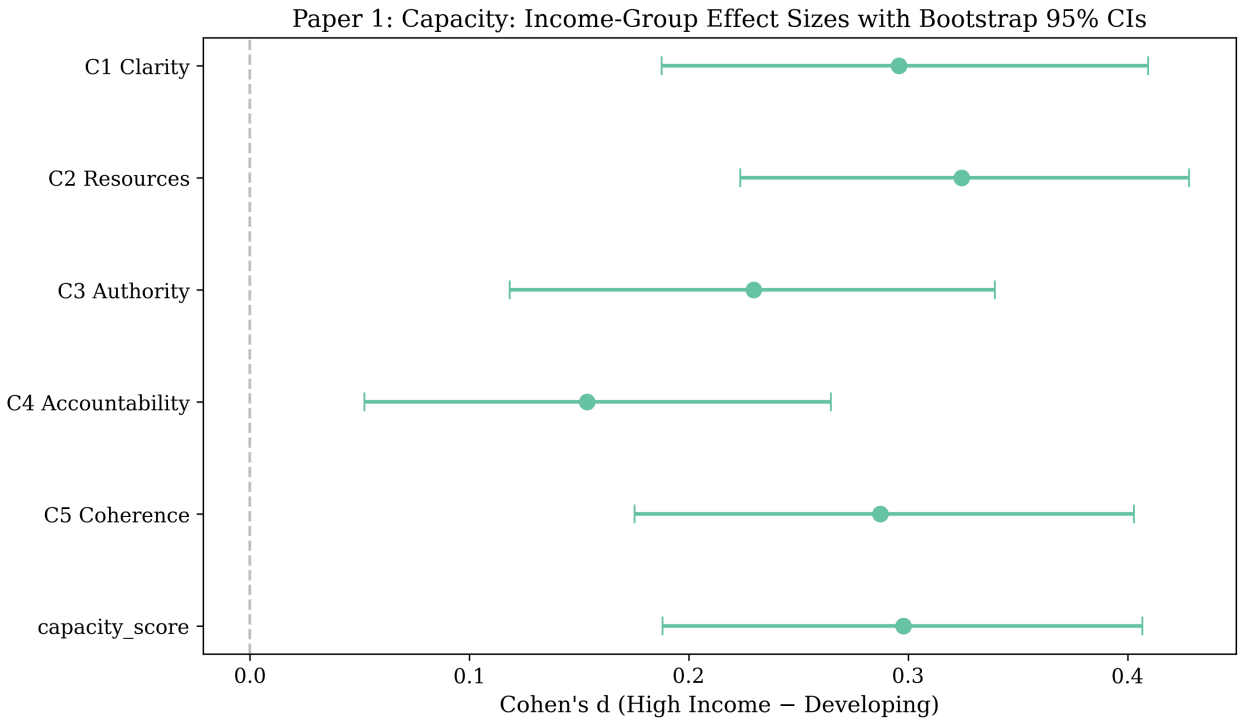

Figure E.1: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for capacity.

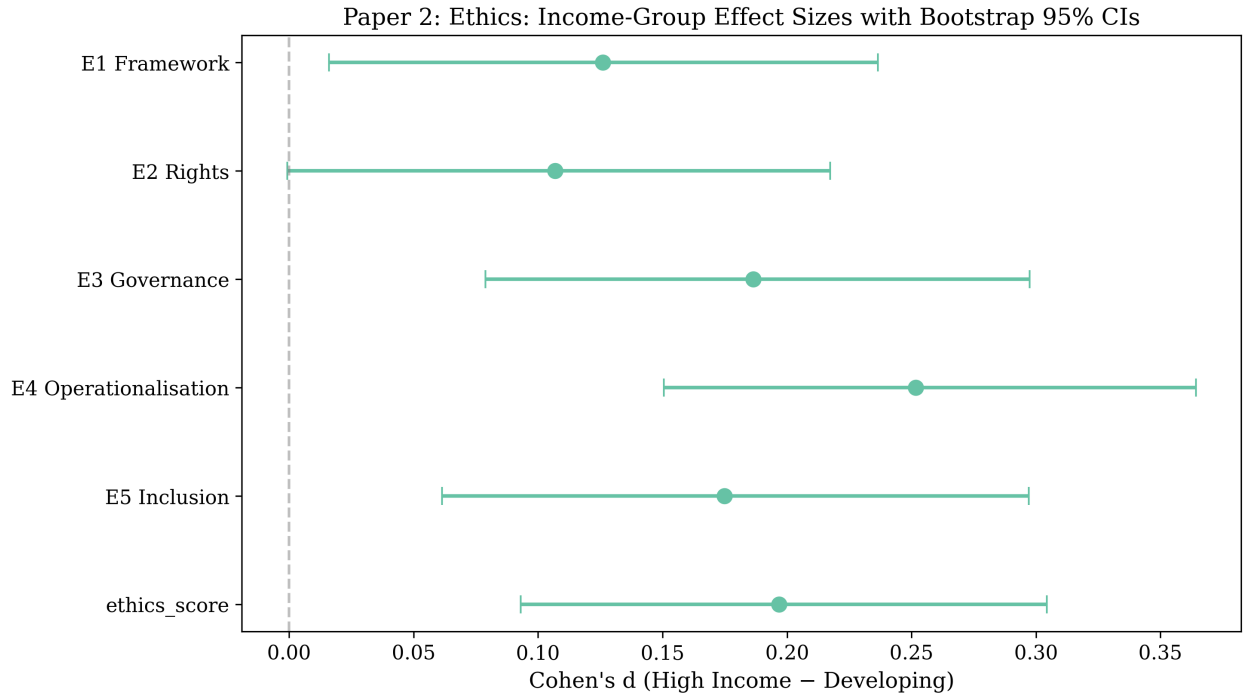

Figure E.2: Bootstrap distributions of the income-group effect size (Cohen's d) from 1,000 resamples for ethics.

E.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 E.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 E.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-

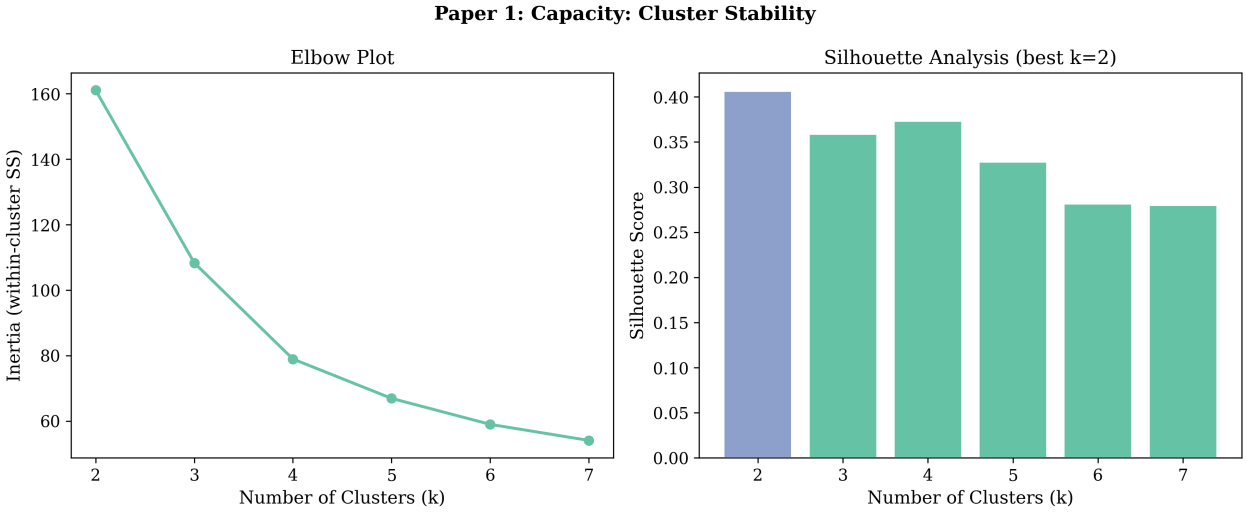

Figure E.3: Cluster stability analysis across different values of k for capacity dimensions.

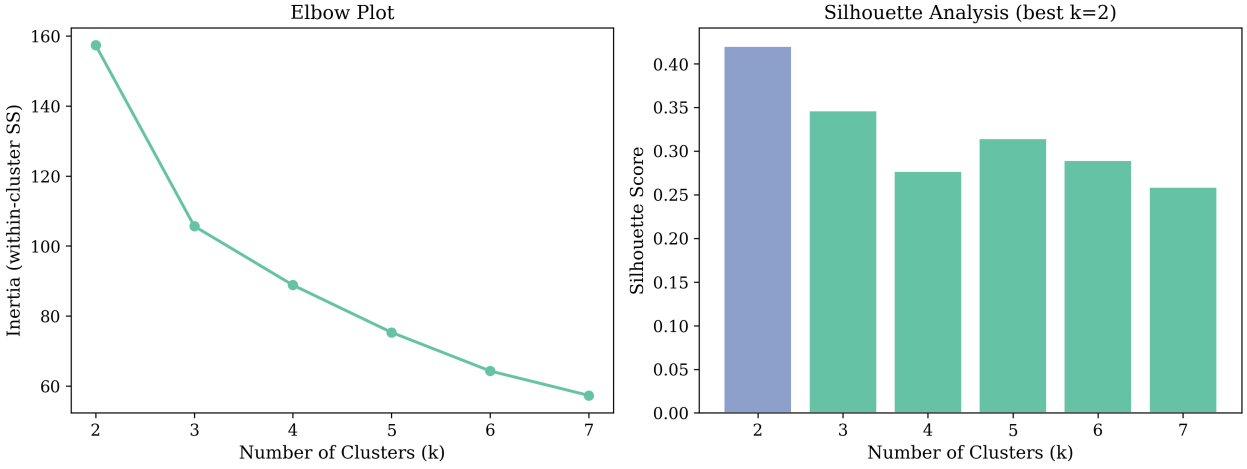

Figure E.4: Cluster stability analysis across different values of k for ethics dimensions.

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.

E.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.

E.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 E.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.

E.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 E.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.

E.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 E.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.

E.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 E.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.

E.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 E.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).

E.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 E.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.

E.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 E.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.

E.1.5 Regression Diagnostics

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

Table E.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.

E.1.6 Multilevel Model Specifications

The multilevel models reported in **?@sec-cap-multilevel** and Section 7.1.3 were estimated using restricted maximum likelihood (REML) with the `lme4` package in R. We report full variance decomposition and model comparison statistics.

Table E.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.

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

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.” Berkman Klein Center, Harvard University.

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.

Hagendorff, Thilo. 2020. “The Ethics of AI Ethics: An Evaluation of Guidelines.” *Minds and Machines* 30: 99–120.

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.”