

1. Introduction

A country registers as well-positioned on every major AI readiness index while its AI sector develops faster than any institution is resourced to govern. The Oxford Insights Government AI Readiness Index (2023), the Global Innovation Index (WIPO, 2024) and the Network Readiness Index (Dutta and Lanvin, 2024) would not flag this. They cannot. Their weighting schemes allow strength in one dimension to offset weakness in another, so the gap between capability and governance never surfaces in the output. It has been averaged away. This is not a technical flaw in any individual index. It is a structural property of compensatory aggregation: when qualitatively different dimensions are made commensurable and summed, the distinction between developing AI capability and governing it dissolves in the arithmetic. These indices ask which countries are best positioned to benefit from AI. They do not ask where AI-driven disruption is outpacing the institutional capacity to manage it. Those are different questions, and the gap between them is where unmanaged labour displacement concentrates. Answering the second question requires an instrument that keeps capability and governance visible as separate, interacting dimensions rather than collapsing them into a single compensatory score.

This dissertation develops and tests a non-compensatory Build-Break-Balance measure of AI governance risk across a 48-country sample. The structure borrows from disaster risk geography, the logic of hazard, exposure and vulnerability (Wisner et al., 2004; Adger, 2006), but adapts it as an organising principle rather than importing the model wholesale. Three dimensions are kept analytically distinct and combined non-compensatorily, so that governance weaknesses cannot be offset by capability strengths.

Build captures where AI capability physically concentrates: the material infrastructure and digital connectivity that make deployment and diffusion possible. Break captures where labour markets face AI-driven task exposure. But what that exposure means depends on who is in the workforce. The same score describes a fundamentally different situation in an economy with a young, expanding labour supply than in one where the working-age population is shrinking. In the first case, automation competes directly with available workers for the tasks they currently perform. In the second, it may be filling gaps that demographic change has already created. The framework decomposes exposure accordingly, distinguishing displacement pressure from shortage mitigation. Balance captures whether governance institutions have the implemented capacity to manage disruption, not whether strategies have been published or principles endorsed.

The framework treats AI governance risk as geographically uneven, produced through the spatial distribution of infrastructure, labour-market demographics and state capacity. Where capability outpaces governance, where exposure meets inadequate protection, where AI

systems developed in one jurisdiction propagate risk into another: these are the configurations the framework is built to identify, and they are exactly what compensatory indices compress into invisibility. It contributes by re-specifying AI capability as infrastructural hazard rather than a readiness asset, conditioning labour exposure through demographic structure to distinguish displacement pressure from shortage mitigation, and measuring implemented governance through an original cross-national coding protocol that distinguishes de jure rules from de facto enforcement.

Three research questions structure the analysis:

1. Which countries combine high AI capability, displacement-oriented labour exposure and weak implemented governance, and therefore register the highest AI governance risk?
2. How does demographic structure alter the interpretation of AI task exposure as labour-displacement risk or labour-shortage mitigation?
3. What distinct Build-Break-Balance configurations emerge across the sample, and to what extent are governance scores spatially clustered?

The analysis is country-level and cross-sectional, covering 48 countries stratified across World Bank income groups and UN geoscheme regions to ensure variance across all three pillars and to avoid constructing a de facto GDP proxy. Governance capacity is scored through an original coding protocol applied to 92 policy documents, with key design choices tested for robustness across multiple specifications. The framework does not capture subnational variation, firm-level governance, or observed employment outcomes. Chapter 2 develops the conceptual foundations for each pillar. Chapter 3 operationalises the framework and describes the governance coding protocol, risk specification, and robustness architecture. Chapter 4 presents the cross-national results. Chapter 5 discusses the findings and their implications.

The dissertation argues that AI governance risk is highest where concentrated capability and displacement-oriented labour exposure meet weak implemented governance, a pattern that readiness rankings flatten or miss.

3.8 Validation and Robustness

Table 1: Robustness architecture

Threat Family	Test	Key Parameter	Scope
Specification sensitivity	OADR removal	Exposure-raw replaces Break-displacement	1 scenario
Specification sensitivity	Sovereignty weighting	0.4/0.6 and 0.6/0.4 variants	2 scenarios
Specification sensitivity	Federal bounds	10%, 15%, 20% population thresholds	3 scenarios
Specification sensitivity	Unknown-as-zero	Unknown governance scores treated as 0	1 scenario
Specification sensitivity	Winsorisation	90th, 95th (baseline), 99th percentiles	2 alternative scenarios
Statistical uncertainty	Monte Carlo simulation	Calibrated noise on all indicators	1,000,000 iterations
Statistical uncertainty	Parameter sweep	Normalisation, weighting, aggregation	100,000+ configurations
Statistical uncertainty	Bootstrap resampling	Cluster stability under resampling	500,000 iterations
Structural influence	Leave-k-out	k = 1 to 4, country removal	213,000+ combinations
Structural influence	Adversarial perturbation	Min. data change for rank/cluster shift	50,000 per country
Structural influence	Shapley decomposition	Indicator contribution to risk	Per country
Structural influence	Permutation tests	Pillar correlations, spatial autocorrelation	Per pillar pair
Structural influence	Rank reversal matrix	Pairwise statistical distinguishability	All country pairs

Robustness assessment targets three distinct threat families (Table 1). The first is specification sensitivity: do the rankings survive alternative design choices? Nine specifications test this directly -- removing the OADR decomposition, varying sovereignty weightings (two variants), adjusting federal bounds (10%, 15%, 20% thresholds), treating unknown governance scores as zero, and varying winsorisation percentiles (90th, 99th against the 95th baseline). Rank correlations and cluster membership changes are reported across all specifications. Findings that do not survive are flagged as fragile.

The second threat is statistical uncertainty: how precise are the rankings given measurement noise? Monte Carlo simulation adds calibrated noise to all indicators simultaneously, producing rank confidence intervals for every country. A parameter sweep varies normalisation method, weighting scheme, and aggregation rule to test whether rankings

reflect substantive differences or methodological choices. Bootstrap resampling tests cluster stability, identifying swing countries whose cluster assignment is sensitive to sample composition.

The third threat is structural influence: which countries and indicators are driving the results? Leave-k-out analysis identifies countries whose removal disproportionately affects others' rankings. Adversarial perturbation finds the minimum data change required to shift each country's rank by 10 positions or flip its cluster assignment. A Shapley value decomposition (Shapley, 1953) attributes each country's risk score to individual indicator contributions. Permutation tests assess whether pillar correlations and spatial autocorrelation are statistically significant. A rank reversal matrix identifies which country-pair rankings are statistically distinguishable and which are indistinguishable under measurement uncertainty. Full specifications are in Appendix A.

Pipeline validation combines internal consistency (the verification architecture ensures no fabricated evidence enters the scoring), stratified output review against source documents spanning different clusters and documentation conditions, and two external checks. The iterative prompt refinement process is documented in the decision log, with the Worker Protection revision representing the most material change and its effect on scoring distributions reported in Chapter 4. *Balance-raw* is compared with World Bank Regulatory Quality (World Bank, 2024b) as a convergent validity test -- the expected relationship is positive but imperfect; too high a correlation would suggest the measure captures general state capacity rather than anything AI-specific. Moran's I (Moran, 1950) provides a descriptive spatial check using queen contiguity weights with nearest-neighbour fallback for island states; significant positive autocorrelation would be consistent with regional governance diffusion, particularly within the EU.

3.9 Limitations and Reproducibility

The most consequential limitation is that *Balance* measures documented governance capacity, not governance capacity directly. States producing extensive regulatory text score higher partly because there is more material to code. The sample includes documents in Arabic, Portuguese, Spanish, French, German, and other languages processed through English translations, which may reduce scoring granularity. The convergent validation partially addresses this, but the tension between documentation and reality cannot be fully resolved.

The demographic decomposition, while theoretically grounded, is empirically dominant in the risk specification. The sensitivity analysis removing OADR produces markedly different rankings, indicating the decomposition is doing substantial rather than marginal analytical work. This should be read as both a strength and a limitation: the decomposition captures real variation in how AI exposure manifests across demographic contexts, but it means the risk

4.5 Uncertainty and Validation

The robustness suite quantifies what the rankings can and cannot support. Monte Carlo simulation (one million iterations) produces rank confidence intervals that are narrow at the extremes and wide in the middle. Japan has a 90% CI width of 2 rank positions; the UAE has a width of 7; Singapore has a width of 6. At the other extreme, Vietnam has a CI width of 42, India 37, Indonesia 35 and Ghana 35, meaning their ranks could plausibly shift across most of the table under measurement uncertainty. The pattern is systematic: the top and bottom quartiles of the ranking have mean CI widths of 13.4 and 13.8 respectively, while the second and third quartiles average 23.1 and 21.8. Countries near the extremes are far from their neighbours in risk space; countries in the middle are congested. Some nominal rank differences are effectively meaningless. Australia (rank 8) and the Philippines (rank 17) are statistically indistinguishable, with a rank-reversal probability of exactly 0.500 despite being separated by nine ordinal positions. Mexico and the Philippines (ranks 15 and 17) are similarly indistinguishable. Thirty-eight country pairs in total have overlap probabilities between 0.45 and 0.55.

Eight of nine sensitivity scenarios produce Spearman rank correlations above 0.95 with the baseline specification. Sovereignty weighting, federal bounds, missing data treatment and winsorisation percentile all change the rankings negligibly. The single fragile scenario is demographic adjustment: removing OADR produces a rank correlation of 0.504 and alters cluster membership for 27% of countries. This fragility is not a deficiency in the robustness architecture. It is a substantive finding about what the framework is measuring: the demographic decomposition is doing real analytical work, and the rankings depend on whether that decomposition is accepted.

The three pillars measure distinct constructs. As reported in Section 4.1, Hazard and *Break-displacement* are not significantly associated, and Hazard and Balance are moderately positively associated. The remaining permutation test confirms that Break and Balance are moderately negatively associated ($\rho = -0.297$, $p = 0.041$), meaning countries with higher displacement exposure tend to have weaker governance. The framework is not collapsing into a single latent factor.

Balance-raw correlates with World Bank Regulatory Quality at Spearman $\rho = 0.566$ ($p < 0.001$). The moderate magnitude is itself informative. A stronger correlation would suggest Balance is simply proxying for general state regulatory capacity; a weaker one would raise questions about whether it measures governance at all. At 0.566, Balance shares enough variance with established governance indicators to confirm it tracks institutional capacity, while remaining distinct enough to support the claim that it captures something specific to AI governance implementation that existing measures do not.

Adversarial perturbation analysis identifies Vietnam, France and Sweden as the most fragile country positions, requiring perturbations of less than 8% of indicator range to shift their ranks by ten positions. Japan, Saudi Arabia and Rwanda are the most robust, requiring perturbations exceeding 23%. That France and Sweden are rank-fragile while belonging to the most stable cluster (European, bootstrap 0.96) is not contradictory: cluster membership is robust under resampling, but ordinal position within a congested region of the distribution is sensitive to small data changes. Leave-k-out influence testing confirms that no single country's removal reshapes the rankings dramatically; the most influential countries are the low-income cluster members (Ghana, Bangladesh, Pakistan), whose removal affects others' relative positions because they anchor one end of the distribution. Full results are reported in Appendix D.

The framework's primary utility is typological. Ordinal precision is not the point. Precise mid-table rankings are not robust, and the framework cannot reliably distinguish between countries in the congested middle of the risk distribution. What is robust across specifications is the identification of structural configurations: the governance gap as a specific pattern concentrated in high-capability, young-workforce economies with weak institutions; the demographic buffer that transforms displacement pressure into shortage mitigation in aging economies; and the five cluster archetypes that map these configurations onto geographically coherent groups. These are the findings the data can support, and they are what the framework was built to see.