MIT AI Risk Initiative

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS OF PROGRESS IN COMPUTING

Massachusetts Institute of Technology

CSET

Mapping the AI Governance Landscape

April 2026 Update

April 2026

Simon Mylius, Peter Slattery, Yan Zhu, Mina Narayanan, Adrian Thinnyun, Suhani Gharia, Daniela Muhaj, Bosco Hung, Victoria Snorovikhina, Alexander Saeri, Jess Graham, Michael Noetel, and Neil Thompson

Executive Summary

What we did

- We improved our LLM-based pipeline to classify over 1000 AI governance documents from the Center for Security and Emerging Technology's AGORA (AI Governance and Regulatory Archive) dataset.
- We now classify risks against six taxonomies: risk domain coverage (from the MIT AI Risk Taxonomy's 24 subdomains), sectors governed, AI lifecycle stages, AI actors, legislative status, and AI system technical scope.
- We refined the methodology by transitioning from a 5-point to a more reliable 3-point coverage scale, updating LLM prompts to reduce overconfidence in scoring, and expanding human evaluation to calibrate classifications.

What we found

- Governance documents within the AGORA dataset concentrate heavily on model safety risks such as security vulnerabilities, privacy, and transparency, whereas socioeconomic risks like economic devaluation, power centralization, and emerging concerns like multi-agent risks and AI-welfare receive comparatively little attention.
- Coverage is uneven across sectors and lifecycle stages: public administration and scientific R&D dominate, while consumer-facing and labor-intensive sectors have lower representation.
- Most lifecycle stages are addressed by a majority of documents, but downstream stages (Deploy, Operate, and Monitor) receive notably greater attention than early-stage data practices.
- Governance framing tends to be broad, targeting 'AI Systems' and 'AI Models' in general terms with limited attention to frontier, foundation, or open-weight systems.
- These patterns point to potential gaps in governance coverage of socioeconomic risks, early lifecycle stages, and consumer-facing sectors, though further analysis is needed to determine where these gaps are most consequential.

What's next

- We welcome feedback and expressions of interest in engaging with our work.
- We plan to integrate this Governance Mapping dataset with other MIT AI Risk Initiative data to build a fuller picture of the AI risk landscape. By linking governance coverage with real-world incidents, known mitigations, and expert assessments of vulnerability and responsibility, we aim to identify not just where governance gaps exist, but where they matter most.

Feedback	Expressions of interest

Contents

License

Suggested citation

Simon Mylius, Peter Slattery, Yan Zhu, Mina Narayanan, Adrian Thinnyun, Suhani Gharia, Daniela Muhaj, Bosco Hung, Victoria Snorovikhina, Alexander Saeri, Jess Graham, Michael Noetel, and Neil Thompson (2026). *Mapping the AI governance landscape: April 2026 update.*

Research Motivation

The rapid evolution of AI governance has outpaced answers to some basic questions: what do governance frameworks actually cover, how comprehensively do they address known risks, and where do critical gaps exist?

To address these questions, we have built a pipeline to map existing AI risk governance documents from the AGORA (AI Governance and Regulatory Archive) dataset, maintained by the Emerging Technology Observatory at the Center for Security and Emerging Technology (CSET). Our pipeline generates reports, visualizations, and a database that enable users to explore the AI governance landscape and understand which AI risks are being addressed by current frameworks across sectors, use cases, and stakeholders.

We discussed our initial pilot in this blog post. We have now expanded our analytical framework to provide more nuanced insights into the governance landscape, including classifications of AI Lifecycle stages, key Actors involved, and Technical Scope. This multi-dimensional approach allows us to identify governance capacity and gaps with greater specificity, understand who is involved in addressing different risks, and assess both the intended scope and topical coverage of governance documents.

Our methodology has also been refined based on lessons from the pilot phase. We have transitioned from a 5-point coverage scale to a more reliable 3-point scale (No coverage, Minimal coverage, Good coverage), which our initial analysis revealed was better suited for assessing coverage across sectors and risk domains. Additionally, we have expanded our evaluation dataset to include a larger sample of human-reviewed classifications and updated our prompts with explicit instructions to address previously identified misclassification patterns.

Methodology Updates

Building on our initial research, and in response to ongoing stakeholder and user feedback on our initial report and research tools, we made the following methodological updates:

Pivot to a 3-point coverage scale

During the pilot stage of this project, we found not only that risk subdomain and sector coverage assessments made by the LLM frequently disagreed with human consensus, but also that two human expert reviewers frequently disagreed with each other. The 5-point scale required raters to draw distinctions such as between “basic” and “minimal” coverage, or between “good” and “excellent”, that were difficult to apply consistently, particularly given the varied language that governance documents in the dataset use to describe sectors and risk domains.

To address this, we consolidated the original 5-point scale into a 3-point scale: "basic" and "minimal" coverage were merged into a single "minimal coverage" category, "good" and "excellent" coverage were merged into "good coverage," and "no coverage" was retained as-is.

Updated prompts to reduce misclassifications

We used an evaluation method based on Quadratic Weighted Cohen's Kappa to assess inter-rater agreement as outlined in our project pilot blog post. We noticed a misclassification pattern on coverage scores where the LLMs frequently assigned a higher coverage score than human expert reviewers. Reviewing the reasoning returned with the LLM responses, we found that coverage was being over-scored when documents included mention of a topic area, even when the specific risk was not covered.

To address this, we updated the prompts with explicit guidance to reduce overconfidence, including examples of a "concept mention" that should not be counted, vs. "risk coverage" that should be counted. We also provided some guidance for the LLM to distinguish between minimal and good coverage based on the length of the content in the documents related to the risk or sector being assessed: "no more than a few sentences" for minimal coverage and "typically at least a paragraph" for good coverage.

Full definitions for the coverage scales that were provided to the LLM are included in Appendix 1.

Taxonomies

We used our LLM pipeline to classify each document in the dataset according to the following 6 taxonomies:

1. **Risk Subdomain Coverage** based on the MIT AI Risk Taxonomy: The level of coverage of each of the **24 subdomains** in the taxonomy was assessed by the LLM for each document in the AGORA dataset, according to the 3-point scale (no coverage, minimal coverage, good coverage).

2. **Sectors Governed**: We used a taxonomy of **14 sectors** based on the North American Industry Classification System. The coverage of each sector was assigned a rating (no coverage, minimal coverage, good coverage) for each document.

3. **AI Actors**:
   We defined 4 different roles in the governance process:
   - **Proposers** (who draft governance instruments)
   - **Targets** (who must comply)
   - **Enforcers** (who oversee compliance)
   - **Monitors** (who track effectiveness)

For each document, the LLM attempted to identify the types of entities fulfilling each role, choosing from the following list of entity types:

- **AI Developer**
- **AI Deployer**
- **AI Governance Actor**
- **AI User**
- **AI Infrastructure Provider**
- **Affected Stakeholders**

Where available, entity names were extracted. Definitions of actor types and roles are provided in Appendix 2.

4. **Stage of the AI Lifecycle**: The LLM identifies which stages of the AI Lifecycle are covered by each document. We use the OECD / NIST definitions of the six lifecycle stages:
   - **Plan and Design**
   - **Collect and Process Data**
   - **Build and Use Model**
   - **Verify and Validate**
   - **Deploy**
   - **Operate and Monitor**

5. **Legislative Status**: The LLM classifies each document as either:
   - **Hard Law**
   - **Soft Law** (non-binding principles, agreements, declarations, guidelines, or standards that rely on voluntary adherence or normative pressure)
   - **Other** (internal corporate policy documents or hybrid, experimental, or emerging governance mechanisms that do not fit traditional hard/soft law categories)

6. **AI System Technical Scope**: A non-exclusive and non-comprehensive list of technical features describing AI systems, including:
   - **AI Models**
   - **AI Systems**
   - **Frontier AI**
   - **General Purpose AI**
   - **Task-specific AI**
   - **Generative AI**
   - **Predictive AI**
   - **Compute Threshold**
   - **Open-weight or Open-source**

Results from Coding the AGORA Dataset

Important context for interpreting these results:

- The findings below are based on the approximately 1000 documents in the AGORA dataset, which is predominantly composed of U.S.-origin English language government documents, the majority of which are federal-level.
- Coverage patterns described below therefore reflect the priorities and framing conventions of this particular corpus and should not be taken as representative of the global AI governance landscape.
- Coverage scores should be taken as indicative of broad patterns rather than precise measurements. We have found LLM classifications to exhibit some biases including over-attribution of coverage when governance-related language is present.

1. Coverage of Risk Domains

Dominance of model safety risks

The risk subdomains covered by the highest number of documents are AI system security vulnerabilities and attacks, Governance failure[1], Lack of capability or robustness, Compromise of privacy, and Lack of transparency or interpretability. In contrast, the least frequently[2] covered subdomains include AI welfare and rights, Multi-agent risks, Economic and cultural devaluation, Power centralization, and Environmental harm.

The most-covered subdomains tend to focus on model safety and established regulatory concerns (security, privacy, transparency), while less-well-covered subdomains include socioeconomic risks (economic devaluation, power centralization) and emerging considerations (multi-agent risks, AI welfare).

[1] See note on Governance Failure in Limitations of the LLM Classification

[2] See 'Fragmentation and Uncertainty: Mapping AI Regulatory Trends and Engagement in the U.S.' (forthcoming) for further discussion

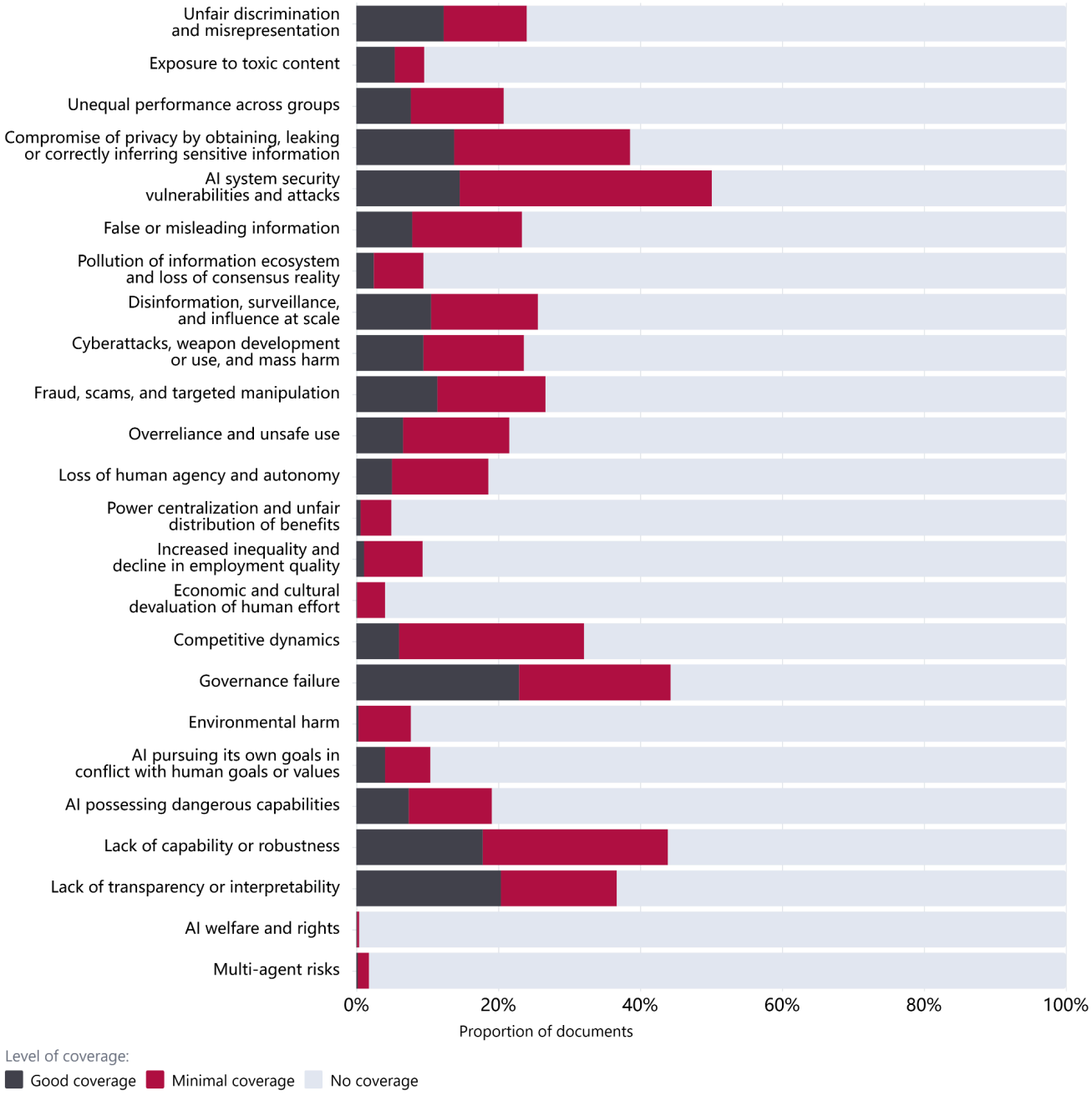

Figure 1: Coverage level of risk subdomains

Policy implication

Our results suggest greater coverage of model safety and other established regulatory concerns within the AGORA corpus and relatively lower coverage of socioeconomic and emerging subdomains. These findings may help guide further diagnostic work to determine whether there are policy-relevant governance gaps in AI governance more broadly.

2. Coverage of Sectors[3]

Concentration in public and research-oriented sectors

The sectors most frequently referenced in the AGORA dataset are concentrated in the public sector and research-oriented domains, including Public Administration (excluding National Security), Scientific Research & Development, Information, and National Security.

Consumer-facing and labor-intensive sectors receive less coverage

Sectors more directly tied to everyday life and economic activity receive substantially less attention. The least frequently covered sectors include Accommodation, Food, and Other Services; Management, Administration, and Support Services; and Real Estate and Rental and Leasing.

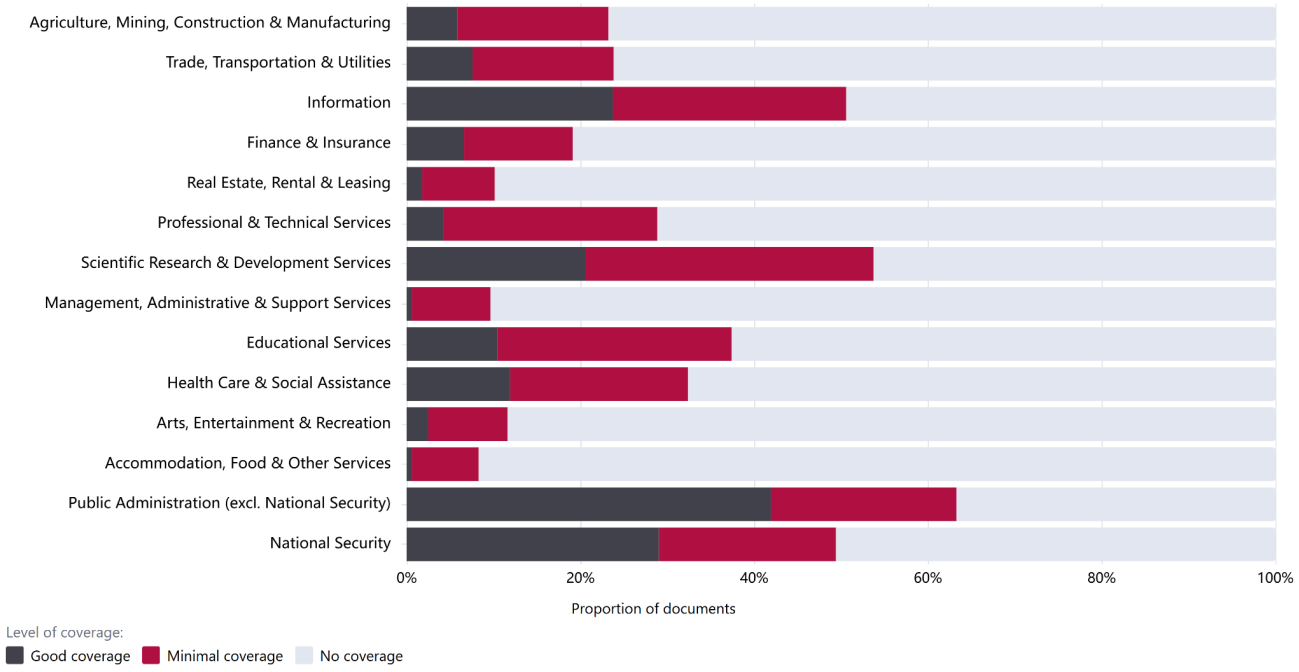

Figure 2: Coverage level of industry sectors

Policy implication

This distribution suggests that AI governance documents within the AGORA dataset have prioritized institutional, technical, and security-oriented contexts, over consumer-facing and labor-intensive sectors. It is unclear whether these differences in coverage are appropriate. It may therefore be valuable to compare current policy coverage against sector-specific vulnerability assessments to monitor for related gaps and oversight.

[3] See 'Fragmentation and Uncertainty: Mapping AI Regulatory Trends and Engagement in the U.S.' (forthcoming) for further discussion

3. Analysis of Actors

Concentration of roles in governance actors

Based on our AGORA dataset, the entity type identified most frequently across all four governance roles (proposer, target, enforcer and monitor) is AI Governance Actor. While all six entity types appear as targets, AI Deployers and AI Developers are each targeted by over 500 documents but have minimal involvement in enforcement and monitoring roles.

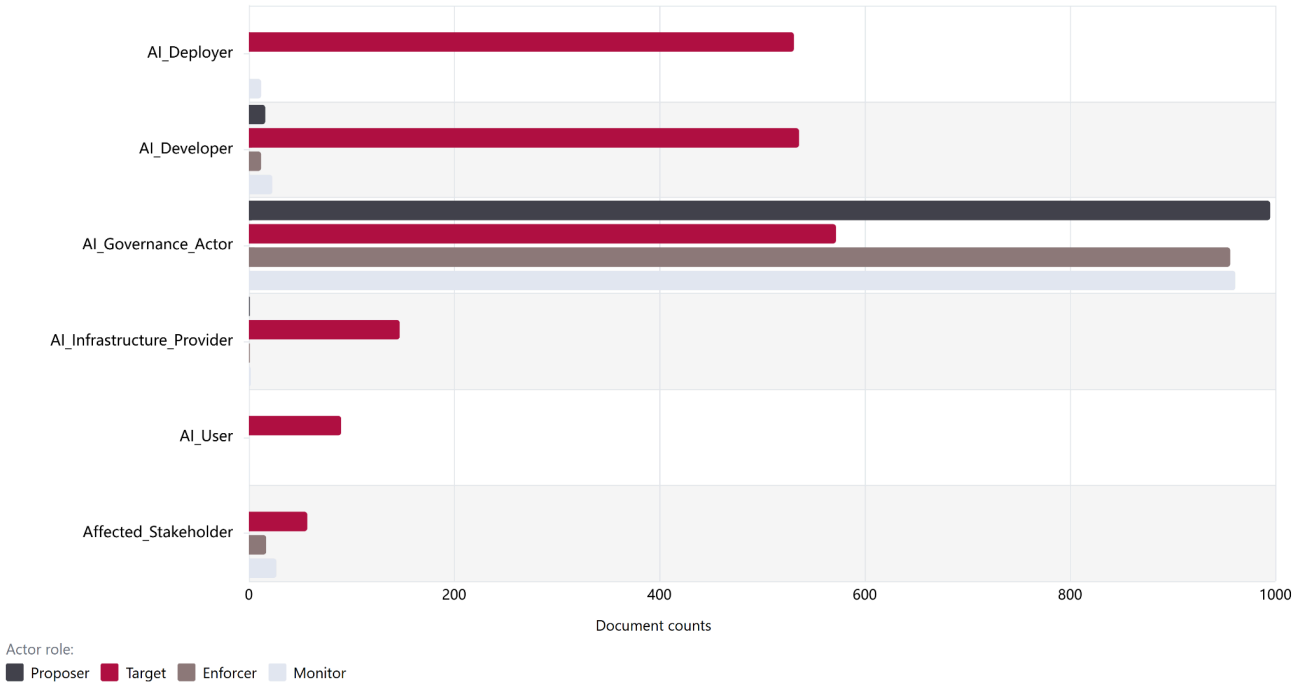

Figure 3: Document counts by actor entity type and role

Policy implication

We find that documents in the AGORA corpus tend to assign formal oversight roles more often to government bodies, while assigning compliance obligations more broadly across developers and deployers. This suggests a broader governance model in which public institutions retain primary responsibility for enforcement and monitoring of AI, even while private actors are expected to bear a substantial share of implementation obligations. Further analysis may be required to assess whether this distribution of governance roles is appropriate or effective.

4. Coverage of AI Lifecycle Stages

Downstream lifecycle stages receive greater attention

Most AI lifecycle stages are covered by more than half of the documents in the dataset, reflecting the fact that many governance documents address multiple stages simultaneously. However, downstream stages (Deploy and Operate and Monitor) are each mentioned by nearly 80% of documents in the AGORA corpus, while Collect and Process Data is covered by roughly half that proportion. This suggests that while governance documents are not exclusively focused on later stages, downstream intervention points receive more attention than early-stage data practices.

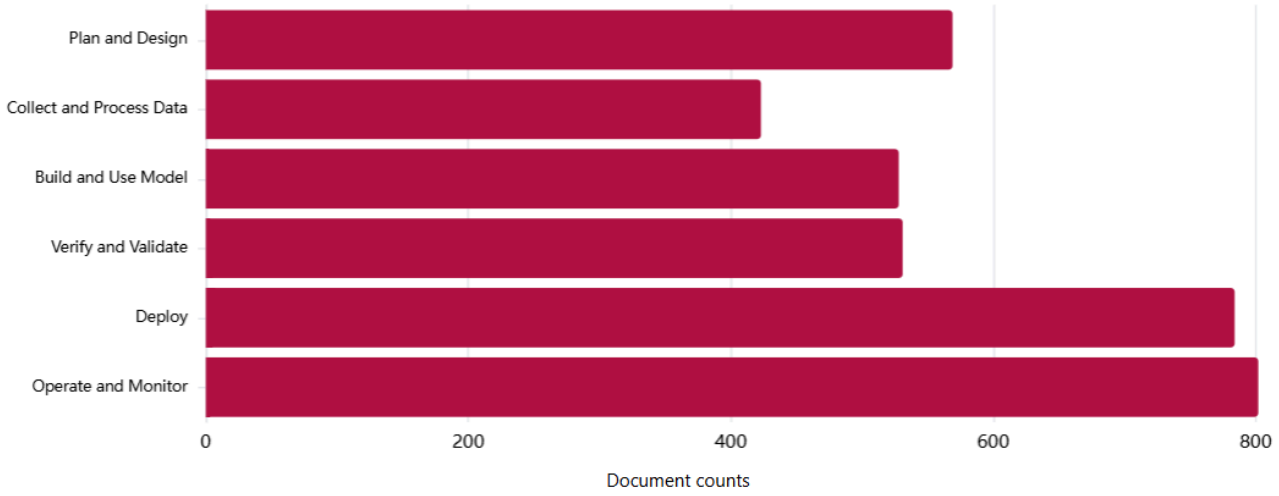

Figure 4: Document counts by AI lifecycle stage

Risks consistently recognized across the lifecycle

A subset of risks that pertain broadly to AI systems are covered across all six AI lifecycle stages. These include AI system security vulnerabilities and attacks, Lack of capability or robustness, Governance failure[4], Compromise of privacy, and Lack of transparency or interpretability.

Policy implication

The relative underrepresentation of early lifecycle stages suggests that governance approaches reflected in the AGORA dataset tend to prioritize downstream controls over early-stage intervention. This warrants further analysis, as it is unclear if the relative differences in coverage are optimal.

[4] See note on Governance Failure in Limitations of the LLM Classification

5. Legislative Status

Hard Law dominates

The overwhelming majority of documents in the dataset are classified as Hard Law, or documents that are legally binding and can be enforced by a court. Of the Hard Law documents in the dataset, only 44% are enacted, with 43% defunct and 12% still proposed - see also "Document validity and timeliness" under Limitations below.

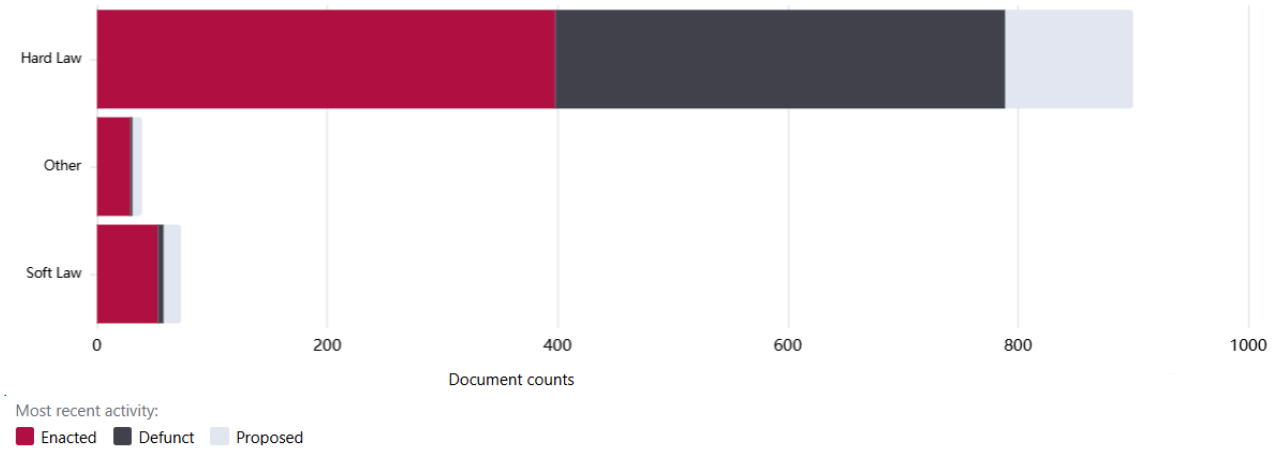

Figure 5: Legislative status of documents split by most recent activity

Policy implication

The high proportion of defunct and proposed hard law documents suggests significant legislative churn. While some degree of turnover is expected given the rapid pace of AI capability development, this pattern may warrant monitoring to understand whether governance efforts are building cumulatively over time or being repeatedly restarted. Future research may be warranted to assess whether legislative churn is higher for documents related to AI governance than in other domains and whether this holds for documents across other jurisdictions.

6. Coverage of AI System Technical Scope

Systems-level governance is most common

Within the AGORA dataset, governance documents most frequently address AI Systems and AI Models. While Generative AI, Task-specific AI, and Predictive AI receive substantial attention, they are referenced less frequently than these broad system-level categories. In contrast, specific terms such as Frontier AI, Foundation Models, Open Weight/Open Source, and Compute Thresholds receive only limited coverage.

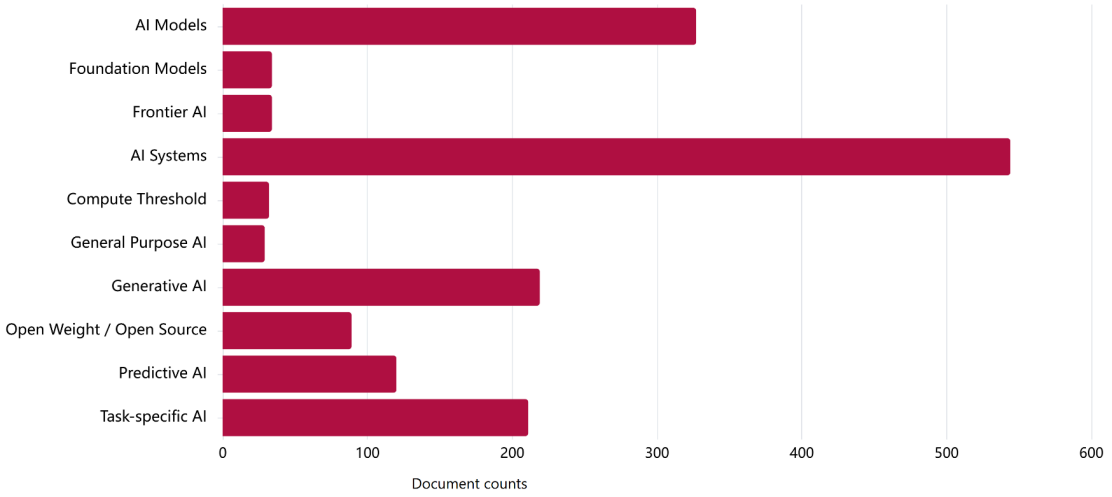

Figure 6: Documents mentioning terms relating to AI technical scope

Policy implication

Our results suggest that governance documents in the AGORA dataset tend to regulate AI in general terms rather than targeting specific system types, particularly those associated with frontier AI risks. Regulating AI in general terms may limit the effectiveness of governance for AI systems with distinct risk profiles, such as open-weight systems, as dual-use potential and downstream liability may not be adequately addressed by provisions relating to generic "AI systems".

Limitations

1. Limitations of AGORA Dataset coverage

While the AGORA dataset provides a valuable foundation for mapping AI governance, it has several limitations that affect the scope and generalizability of our findings:

- **Jurisdictional imbalance:** Most of the documents included originate from the United States. This heavy U.S. weighting means that our analysis for the whole AGORA dataset reflects U.S. governance trends more than those across the global landscape. Furthermore, the majority of US documents are federal rather than state or local level legislation.
- **Language constraints:** All documents are provided in English. This prevents us from fully testing how well LLMs interpret governance texts in other languages. This is an important gap, given that policy language is often nuanced and culturally specific, and many non-English documents lack official translations. For some of these cases, AGORA relies on third-party translations, which may introduce inaccuracies. As we expand coverage to more data sources, reliance on unofficial translations, or in some cases the complete absence of translations, may pose further challenges.

AGORA will address some of these challenges by including more CSET translations of non-English documents in the future.

- **Document validity and timeliness:** Not all documents in AGORA are currently in force or reflect the most up-to-date versions. Some laws or policies have expired or been superseded. This temporal mismatch means that certain coverage scores may not perfectly align with the governance frameworks currently shaping AI practice.

2. Limitations of the LLM classification

We conducted follow-up human spot checks on a subset of 20 documents. The results suggest that divergences between LLM outputs and human reviewer judgments may not be randomly distributed within this sample, but instead cluster around specific taxonomies and subdomains. These patterns point to recurring interpretive tendencies in how LLMs map governance language to classification categories, rather than isolated or idiosyncratic errors.

The LLM tended to apply AI lifecycle stages more broadly than human reviewers. In particular, it struggled to consistently distinguish between closely related stages such as *Deploy* vs. *Operate and Monitor*. In addition, the model often inferred lifecycle applicability from downstream implications or anticipated operational effects, even when a document did not explicitly state coverage of those stages. Human reviewers typically only assigned lifecycle stages to a document if it included explicit scope statements or concrete obligations tied to a given lifecycle stage. As a result, lifecycle coverage scores, especially for adjacent or downstream stages, should be interpreted with caution.

The LLM consistently over-attributed coverage of "*Governance Failure*" relative to human consensus. This appears to stem from a tendency to treat the presence of governance-related language, such as references to "risk management," "oversight," or "accountability", as evidence that governance failure is substantively addressed. Human reviewers applied a narrower interpretation, reserving this subdomain for cases where documents explicitly diagnose governance breakdowns or specify mechanisms designed to prevent such failures. Similar, though less pronounced, patterns were observed for subdomains such as *Lack of Transparency or Interpretability* and *AI System Security Vulnerabilities and Attacks*, where high-level principles or generic safeguards were sometimes scored more generously by the LLM than by human reviewers.

The LLM exhibited inconsistent patterns in assigning sector coverage, with both over- and under-attribution depending on how explicitly sectors were named. When government agencies or sector-linked entities were explicitly mentioned (e.g., "Secretary of Agriculture" or "Department of Labor"), the model often inferred sectoral coverage even when those sectors were not the primary focus of the document. Conversely, when sector relevance was implicit, such as a research fellowship program implying *Educational Services*, or references to media literacy curricula implying the *Information* sector, the LLM frequently failed to assign coverage unless explicit sector terminology was used. This suggests that

sector classifications are sensitive to drafting conventions and explicit labeling, rather than solely to substantive intent.

We are addressing these limitations through iterative prompt refinement, clearer operationalization of taxonomy definitions, and additional human calibration. In the interim, the results should be understood as reflecting patterns in how governance documents are framed and signaled, rather than as a definitive assessment of their substantive scope or effectiveness. Ongoing analysis will examine whether these patterns persist beyond the documents included in the current spot checks.

Feedback

We welcome feedback and expressions of interest in engaging with our work.

Below is a summary of the feedback we have received and how it is shaping the next stage of the project:

Across the feedback received, the most consistent theme was a need for greater contextual framing around the visualizations. Respondents noted that they were interpreting graphs without sufficient explanation of what the data represents, how categories are defined, or what conclusions can reasonably be drawn. Specific gaps included undefined terms like "minimal" and "good" coverage, unclear category labels, and an absence of "so what?" framing that would help readers understand the significance of what they are seeing. Several respondents suggested adding inline annotations or brief summaries directly alongside each graph to reduce the risk of misinterpretation before it occurs.

Another prominent theme of feedback focused on the risk that certain visualizations may be misleading without proper context. Reviewers cautioned that document count and frequency-based graphs could be read as progress indicators when they are not, since high document volume does not necessarily reflect effective governance, maturity, or real-world adoption.

Several respondents called for broader structural and functional improvements. These included a dedicated conclusions page, consistent methodology references across all tabs, and clearer upfront definitions. Expanded filtering was widely requested, along with the ability to navigate by Risk Domain.

Acknowledgments

We want to thank the following people for useful contributions and feedback:

- Catherine Barrett
- Tomas Bueno Momcilovic
- Abir Dey

- Tess Hilson-Greener
- Win Myat Nwe Khine
- Sukanya Konatam
- Burak Piskin

Appendix 1: Definitions for Coverage Scores

The definitions below are provided to the LLM as part of the classification prompt:

Risk Subdomain

1. **No Coverage**: Risk subdomain not mentioned at all.
2. **Minimal Coverage**: Brief mention (no more than a few sentences focused on this subdomain). No specific mitigations, recommendations or governance measures relating to the risk domain are described.
3. **Good Coverage**: Comprehensive governance measures or mitigations with clear procedures relating to the risk subdomain are explicitly described in the document (typically at least a paragraph focused on this subdomain).

Sectors Covered

1. **No Mention**: The sector is not mentioned or referenced in any way.
2. **Minimal Coverage**: The sector is mentioned, but there is little elaboration on how AI is governed within it. (no more than a few sentences explicitly relating to the sector being assessed).
3. **Good Coverage**: There is an explicit clear description of governance measures, controls, or obligations relevant to the sector. This would typically be at least several sentences dedicated to this sector.

Appendix 2: Definitions of Actor Types and Roles

Actor Types

- **AI developer:** Entity that creates general-purpose foundation models or specialized AI systems.
- **AI deployer:** Entity that implements AI systems in products/services/platforms used within an organization (internal deployment) or within products/services/platforms delivered to customers or the public (external deployment)
- **AI Governance Actor:** Entity that creates or enforces laws, regulations, standards or guidelines for AI development, deployment, and use
- **AI User:** Entity that uses or relies on AI systems without significant modification
- **AI Infrastructure Provider:** Entity that provides compute, cloud infrastructure, and/or data to train and run AI
- **Affected Stakeholder:** Entity indirectly affected by AI decisions or outputs

Actor Roles

- **Proposer:** Actor who initiates or drafts AI-governance instruments that enter the policy pipeline.
- **Target:** Entity that the governance instrument applies to, regulates, or affects
- **Enforcer:** Authority with mandate to oversee compliance, license, inspect, and sanction
- **Monitor:** Actor who tracks, audits, and reports on implementation and effectiveness of AI governance

Appendix 3: Dissemination and Intellectual Property Approach

1. All deliverables will be designed to serve both academic and practitioner audiences.
2. All deliverables will be hosted and promoted across CSET and the MIT AI Risk Index website.
3. All outputs will be published under open access terms (e.g., Creative Commons Attribution license CC BY 4.0).
4. Acknowledgment of contributions will be made to all team members, advisors, and their organizations.
5. Data collected and coded will be structured for potential reuse in future research.