MIT AI Risk Initiative

FutureTech
THE ECONOMIC AND TECHNICAL FOUNDATIONS OF PROGRESS IN COMPUTING

Massachusetts Institute of Technology

CSET

Mapping the AI Governance Landscape

Pilot test and update

October 2025

Simon Mylius, Peter Slattery, Yan Zhu, Mina Narayanan, Adrian Thinnyun, Alexander Saeri, Jess Graham, Michael Noetel, and Neil Thompson

Executive Summary

What we did

- We tested an approach for using LLMs to categorize legal and governance documents from the Center for Security and Emerging Technology's ETO AGORA (AI GOvernance and Regulatory Archive).
- To develop and validate our approach, we compared the performance of five LLMs and human reviewers at classifying six documents.
- We then used Claude Sonnet 4.5 to classify more than 950 governance documents according to several taxonomies covering AI risks, mitigations, and related governance concepts.

What we found

- Our analysis suggested that, on average, for the six documents we examined, Claude Sonnet 4.5, Claude Opus 4.1, and GPT-5 achieved comparable or greater agreement with human consensus than the agreement achieved between classifications made independently by two human reviewers.
- Provisional findings:
    - The most covered risk subdomains were *Governance failure*, *AI system security vulnerabilities & attacks* and *Lack of transparency or interpretability*.
    - The least covered were *AI Welfare and Rights*, *Multi-agent risks,* and *Economic and cultural devaluation of human effort*.
    - The sectors with most coverage were *Public Administration (excluding National Security)*, *Scientific R&D*, and *National Security.*
    - The least covered were *Accommodation, Food, and Other Services*, *Arts, Entertainment, and Recreation* and *Real Estate and Rental and Leasing*.

What's next

- We welcome feedback and expressions of interest in engaging with our work.
- We will create reports, visualizations, and a database to help users explore the AI governance landscape and understand which AI risks and mitigations are addressed or neglected by current AI governance approaches.

Feedback	**Expressions of interest**

Contents

License

Suggested citation

Mylius, S., Slattery, P., Zhu, Y., Narayanan, M., Thinnyun, A., Saeri, A. K., Graham, J., Noetel, M., & Thompson, N. (2025). *Mapping the AI governance landscape: Pilot test and update.*

Research Motivation

AI governance frameworks are multiplying rapidly. Many governments, standards bodies, and companies have released guidelines, principles, and regulations. However, a basic question remains underexplored: what do these frameworks actually cover, and how comprehensively?

To address this, we are building a pipeline to map existing AI risk governance documents to the MIT AI Risk Taxonomy and other relevant taxonomies. We aim to create reports, visualizations, and a database to help users explore the AI governance landscape and understand which AI risks are addressed and which are neglected by current AI governance approaches.

In the first stage of this work, we focus on documents from the Center for Security and Emerging Technology's ETO AGORA (AI GOvernance and Regulatory Archive), "a living collection of AI-relevant laws, regulations, standards, and other governance documents." The AGORA dataset currently contains over 950 documents and is growing as new documents are added. If successful, we intend to scale up the pipeline to include other data sources and share all outputs under a Creative Commons license (see Appendix 3).

This document describes the methodology we used to pilot our initial approach and some of our initial findings and feedback.

Methodology

After identifying and creating taxonomies for classification (see Appendix 1), we selected six AI-related governance documents from different stakeholders across several U.S. jurisdictions (see Appendix 2) and performed document-level classification and coverage scoring across three layers:

1. **AI Risks** — we labeled the risks from the MIT AI Risk Taxonomy named by each document and assigned **coverage scores (1–5),** which represent the degree to which each risk is addressed in a document**.**
2. **AI Mitigations** — we labeled the mitigations from the preliminary MIT AI Risk Mitigation Taxonomy and assigned **coverage scores (1–5)** representing the degree to which each mitigation is addressed in a document.
3. **Other AI Governance Dimensions** — we labeled each document based on a portfolio of other dimensions derived from several additional **taxonomies** (see Appendix 1).

Six researchers independently reviewed one or more of our six pilot documents and applied our full suite of taxonomies. For every identified AI risk or mitigation measure, each researcher assigned a coverage score from 1 (not mentioned) to 5 (comprehensively addressed). Discrepancies were discussed and reconciled to ensure inter-rater reliability.

In parallel, we used a large language model to analyze each document, generating labels and coverage scores. The model saved direct quotes from the source as evidence supporting each classification. We then compared the performance of human and LLM coder approaches using Cohen's Kappa(κ), a commonly used metric to assess inter-rater reliability.

For classifications based on a 5-point ordinal scale, such as how well a document covers each risk subdomain, we used a quadratic weighted kappa. This metric captured the higher increased significance of disagreement where the difference between the human and LLM coder scores was larger, versus smaller disagreements.

LLM Classification Evaluation

The following are some findings from several rounds of testing LLM classification using the MIT AI Risk domains taxonomy on six documents (see Appendix 2).

General observations

- The LLM analysis initially tended to assign higher scores than the human consensus. We are addressing this by updating the part of the LLM prompt that provides coverage score definitions to include quantification of the different coverage levels in the grading rubric. For instance, for the coverage score definition "2 = Minimal Coverage - Brief mention", we have added the explanatory quantification: "(no more than 1 sentence focused on this subdomain)". This update to the prompt improved the agreement between the LLM classification scores and human consensus by 11% (using weighted Cohen's κ).
- When we compared LLM and human coding, we found that LLMs were often better than humans at:
  - Identifying cases that include concepts discussed in unexpected places. For example, most human reviewers overlooked a mention of *risks to the environment* in the 'Security and Safety' guiding principle of the San José AI Policy, a concept which was picked up consistently over multiple classification runs by LLMs; and
  - Identifying additional risks and mitigations not explicitly mentioned in our categories (e.g., 'other Misinformation risks').

Classification model comparison

- We evaluated the performance of 5 different LLMs in assigning Risk Subdomain coverage scores for the 6 documents in our pilot sample.
- The Risk Domain taxonomy contains 24 subdomains. Our dataset therefore

comprised 24 coverage scores in each of the 6 documents: a total of 144 data points for analysis with each model.

- The models evaluated were: Claude Sonnet 4, Claude Opus 4.1, Gemini 2.5 Pro, GPT-5 and Claude Sonnet 4.5.
- We calculated Cohen's κ scores to assess agreement between the LLM classifications and human consensus. We compared these κ scores with the κ score for agreement between the two human reviewers (H1/H2).

	H1/H2 Human reviewers	Claude Sonnet 4/ Human consensus	Claude Opus 4.1/ Human consensus	Gemini 2.5 Pro/ Human consensus	GPT-5/ Human consensus	Claude Sonnet 4.5/ Human consensus
Mean Cohen's κ	0.764	0.712	0.771	0.700	0.792	0.741
Minimum Cohen's κ	0.396	0.552	0.635	0.365	0.552	0.594
Maximum Cohen's κ	0.927	0.979	0.938	0.958	0.969	0.979

Table 1: Mean, minimum and maximum Cohen's κ scores for the 6 documents evaluated in the pilot stage, showing agreement between 2 human reviewers (H1/H2) and between each of the LLMs evaluated and human consensus.

- Magnitude guidance for Cohen's Kappa from Landis and Koch (1977) characterizes κ values between 0.61-0.8 as indicating "substantial agreement". All 5 models showed a mean Cohen's κ score in this range across the 6 documents assessed.
- The mean κ score for Claude Opus 4.1 and GPT-5 exceeds the κ score for the H1/H2 reviewers, indicating that these models on average would have greater agreement with the human reviewers' consensus[1] than the agreement between the 2 human reviewers' independent classifications.
- The minimum agreement score is an important baseline for the confidence that users should have in the accuracy of our dataset. Looking at the minimum κ score of the 6 documents assessed, Gemini 2.5 Pro showed poorer agreement with human consensus than H1/H2, both within the 'Fair agreement' range as defined by Landis & Koch (1977). GPT-5, Sonnet 4 and Sonnet 4.5 all had a score exceeding H1/H2's, and within the 'moderate agreement' range. The κ score for Claude Opus 4.1 was higher still, within the 'substantial agreement' range.

[1] The consensus classification is the result of a discussion between the 2 human reviewers after they have each independently coded a document.

Preliminary Results from Coding AGORA

We ran the full dataset of over 950 AGORA documents through the data pipeline using Claude Sonnet 4.5 for document classification. We selected Claude Sonnet 4.5 for the classification because it performed well in our pilot evaluation (Table 1 above): it had a higher minimum κ score than GPT-5, and was more economical to run than Opus 4.1 (10% of the cost). An interactive version of these graphs, supporting click-through to view the documents classified in each category, can be explored here.

Coverage of risk subdomains

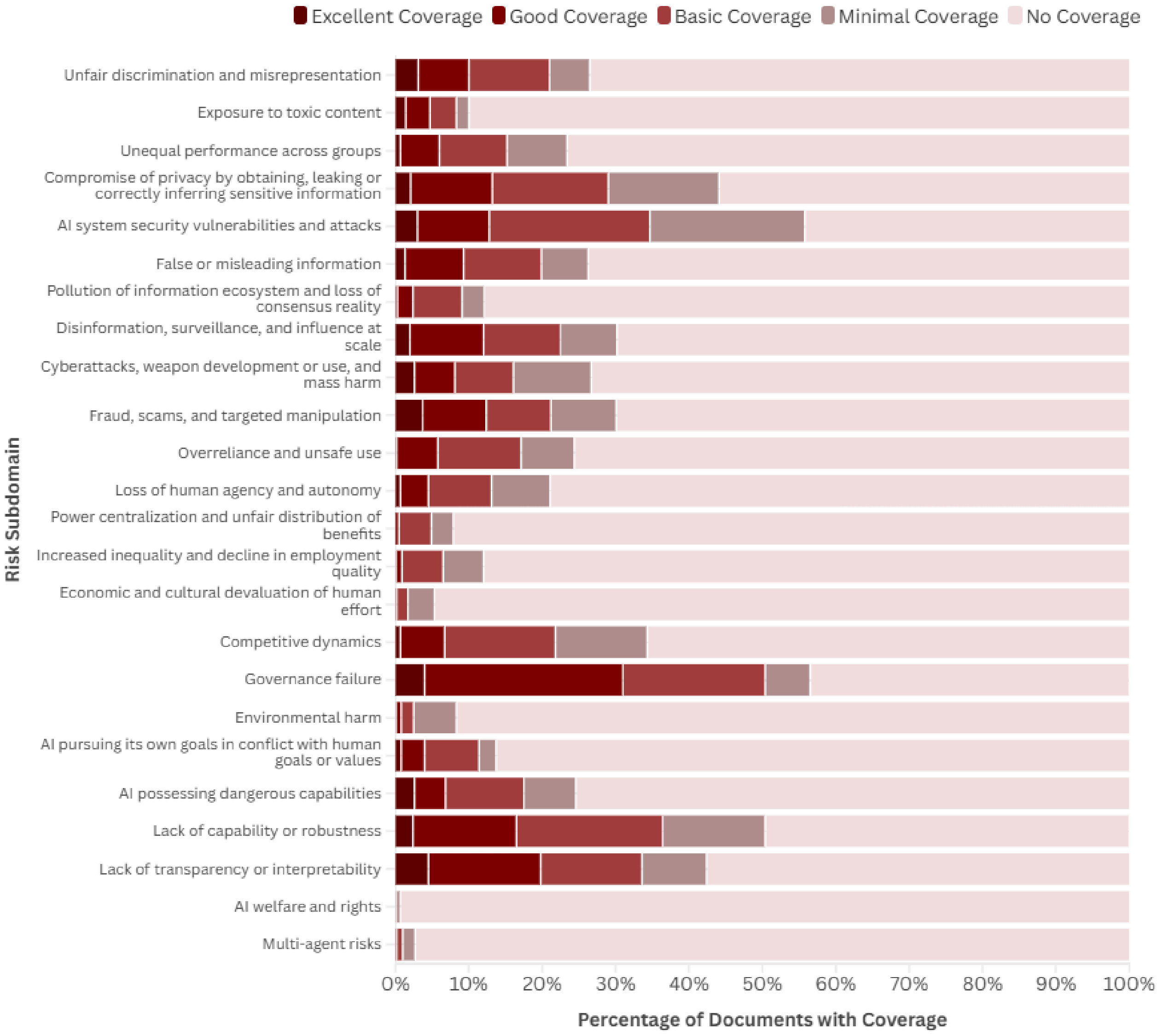

Figure 1: Coverage of risk subdomains by documents in the AGORA dataset

Our preliminary analysis found that the risk subdomains covered by the largest number of documents are *Governance failure, AI system security vulnerabilities and attacks,* and *Lack of capability or robustness.*

As discussed in the 'Limitations of the LLM Classification' section below, we noticed some misclassification of *Governance failure* and will be working to address this, so the shape of this distribution may change.

The risk subdomain covered by the fewest documents in the dataset is *AI Welfare and Rights*, followed by the *Multi-agent risks* subdomain. Most documents released in 2024 or earlier do not cover this risk, reflecting governance's tendency to lag real-world risk emergence.

Coverage of Industrial Sectors

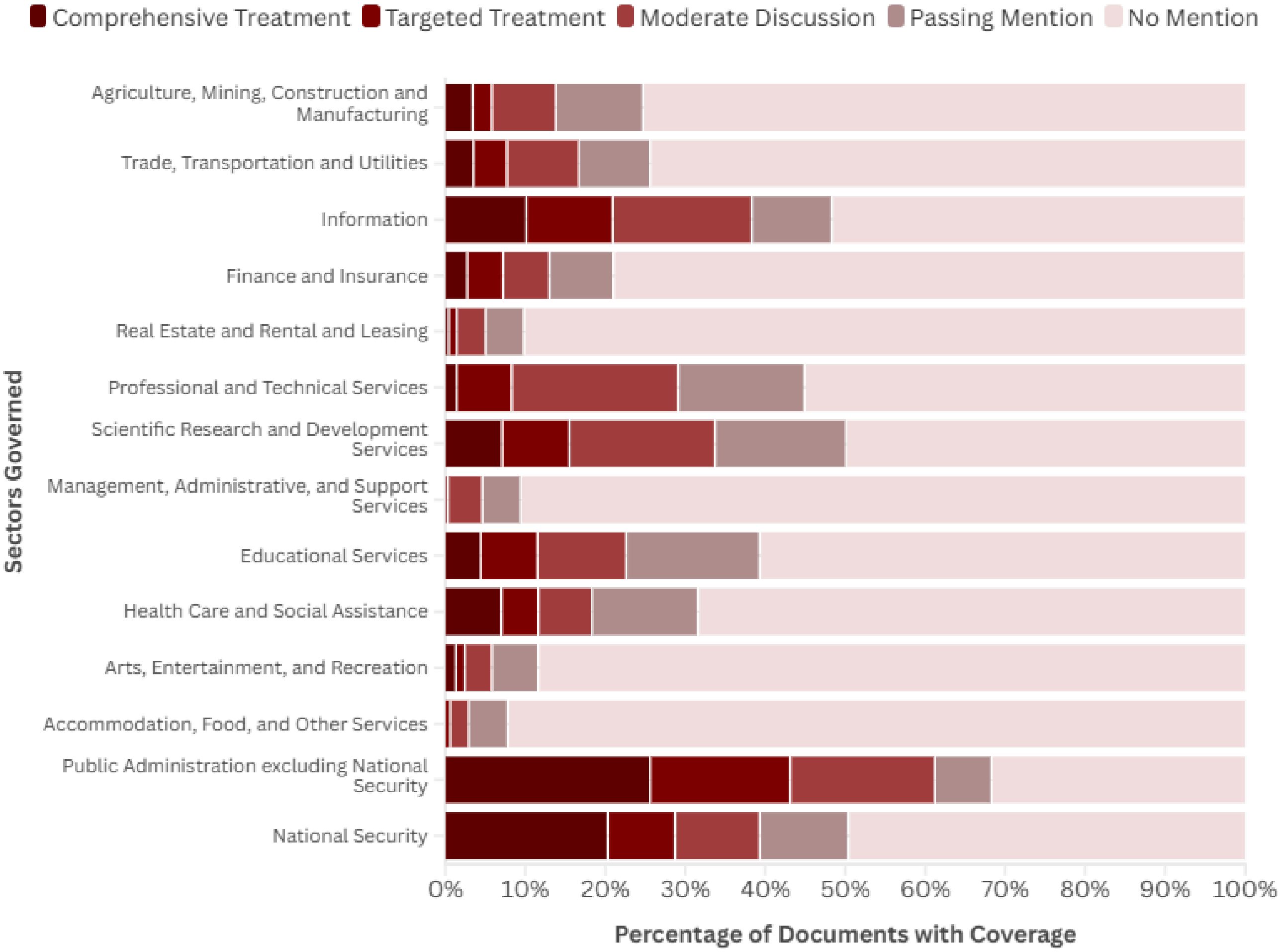

Figure 2: Coverage of industrial sectors by documents in the AGORA dataset

The sectors with the most coverage were *Public Administration (excluding National Security), Scientific R&D*, and *National Security.* The least covered were

Accommodation, Food, and Other Services, *Arts, Entertainment, and Recreation* and *Real Estate and Rental and Leasing*.

Limitations

The following are limitations related to our data source and coding approach, which are important to consider when evaluating our outputs and future work.

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.
- **Language constraints:** All documents are provided in English. This prevents us from fully testing how well LLMs interpret governance texts in other languages—an important gap, given that policy language is often nuanced and culturally specific. Moreover, many non-English documents lack official translations. For some of these cases, AGORA relies on third-party translations, which may potentially introduce inaccuracies. For instance, China's *Measures for the Management of Generative AI Services* was translated by China Law Translate, a Yale Law School–affiliated initiative. 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.

Limitations of the LLM classification

- We reviewed the direct quotes cited as supporting evidence by the LLM for classifications in this category. In several cases, it appears the LLM identifies some phrasing that matches a part of the subdomain/its description but misses the semantic nuance. For example, Claude Sonnet 4 assigned a score of 4 to the San Jose AI Policy (doc 2047) for its coverage of risk subdomain *Governance Failure*

quoting as evidence "Establish and maintain processes to assess and manage risks presented by AI". This quote does indeed relate to governance, but not specifically to the *failure* of governance.
We are addressing this type of misclassification by updating the prompt with explicit instructions to focus on the details of the subdomain descriptions in the taxonomy.

- A 5 point scale may be too granular for assessing coverage of sectors governed and risk domains. We may trial a 3 point scale in the future (e.g., 1 No coverage, 2 Brief mention, and 3 Detailed coverage).
- The LLMs had a bias towards overstating coverage in cases where they were medium confidence. Our prompt instructs the model to assign a (*low*/*medium*/*high*) confidence rating to each subdomain coverage score based on the LLM's assessment of the strength of supporting evidence. We noticed that the subdomains where the confidence rating was *medium* were frequently assigning higher coverage scores than the human reviewers' consensus, whereas subdomains with *high* confidence did not show this tendency. (No subdomains in our sample dataset were assigned *low* confidence.) We are exploring different approaches to address this discrepancy.
- LLM performance differed significantly across documents. The document with the highest κ score (as classified by human reviewers and all 4 LLMs) was the Executive Order on Removing Barriers To American Leadership In Artificial Intelligence (document #1780), with scores in the range 'Almost perfect agreement'. This can be explained by the very low numbers of subdomains covered by this document: human consensus only identified 3 subdomains with coverage scores >1, indicating that the other 21 subdomains were not mentioned at all by the document. The LLMs perform well at the binary 'mention/no-mention' classification.

Limitations of Quadratic Weighted Cohen's Kappa

- The Landis & Koch (1977) magnitude guidance we used (e.g. a κ score in the range 0.61-0.80 = "Substantial agreement") has been acknowledged as being subjective by the authors themselves[2].
- While common, our use of quadratic rather than linear (or other) weighting for differences on the 5-point ordinal scale is also arbitrary.
- A kappa score may indicate very good agreement whilst overlooking systemic misclassifications. For example, if 1 subdomain is consistently misclassified across all documents, the κ score would not point to that and may still indicate 'near-perfect' agreement if most of the other subdomains agree well. It is therefore

[2] They say: 'although these divisions are clearly arbitrary, they do provide useful benchmarks for the discussion.'

necessary to apply wider checks across the set of sampled documents to identify systemic patterns that could be caused by biases or miscomprehensions.

Feedback

We welcome feedback, and expressions of interest in engaging with our work. To avoid redundancy, we are sharing below a summary of the feedback we have already received, which we are using to shape the direction of the next stage of the project:

- **Clarity on scope:** Many governance documents do not clearly specify which AI systems or models they cover (e.g., whether the scope applies to general-purpose AI, systems above a certain FLOP threshold, or all AI models). This ambiguity often creates confusion for both practitioners and reviewers.
- **Taxonomy challenges:** Given that AI governance can be a nebulous concept, some existing taxonomies remain ambiguous despite long descriptions, making it difficult for human reviewers to reach consensus and increasing the risk of systematic bias. For example, the distinction between safety decision frameworks and governance frameworks is not always clear-cut; the scope of "incident reporting" also is ambiguous.
- **Consistent challenges:** Reviewers highlighted difficulties in ensuring scoring consistency across documents. Variations in individual interpretation may lead to differing levels of generosity or strictness when assessing the same taxonomy item, and even a single reviewer may find it challenging to maintain uniform criteria across all evaluations. Such inconsistencies can affect the reliability of comparative results, underscoring the importance of clearer scoring guidance and systematic calibration procedures.
- **Focus on critical domains:** Several reviewers emphasized the importance of highlighting coverage of potentially higher-stakes topics such as frontier AI, catastrophic risks (e.g., loss of control, AI-enabled biothreats, authoritarian lock-in), as well as areas that merit further investigation such as widely discussed risks or mitigations.
- **Additional lenses:** Suggestions include expanding the taxonomies to incorporate dimensions such as jurisdictional comparisons, cooperation levels (unilateral, bilateral, multilateral), etc.
- **Complexity of governance documents:** Some documents pertain to multiple actors, which complicates classification. For instance, California SB 53, CalCompute and Whistleblower Protections (version of Feb, 2025) combines two distinct sections: the first half targeting the public sector, and the second half targeting AI companies. When both sections mention a particular risk or mitigation, they may warrant different relevance scores, creating challenges for both human and LLM reviewers.

- **Potential audiences:** Although our research is still at a pilot stage and our methods require further refinement, feedback suggests that the outputs could already be useful to a wide range of audiences, including:
    - Policy officials and civil society actors conducting or practicing AI governance;
    - Academics and researchers studying AI governance frameworks;
    - Corporate compliance teams that need to navigate AI obligations;
    - Consultancies and legal advisors working on AI strategy; and
    - Members of the public who are interested in AI risk and AI governance.

Expected Outputs

Our final analysis will map each of the over 950 documents in the full AGORA dataset across the following dimensions:

- **Risk domain coverage**: A coverage score (1-5) for each risk subdomain in the MIT risk taxonomy, based on whether the subdomain is mentioned, and if so, the level of detail in procedures, roles, and guidance for implementing specified measures.
- **Mitigation/control coverage**: A coverage score for each type of mitigation from the MIT taxonomy.
- **Legislative status**: A classification of whether the document covers hard law, soft law or anything else.
- **AI lifecycle stage**: A classification of the stages of the AI lifecycle (based on OECD/NIST definitions) covered by the document, and the types of AI system that it applies to.
- **Actors**: The types of entities (AI developer, deployer, governance actor, user, infrastructure provider, affected stakeholder) associated with each role (proposer, target, enforcer, monitor) as well as the names of any actors that are explicitly mentioned.
- **Industrial sectors governed**: A classification based on the North American Industry Classification System
- **Additional metadata** from the AGORA dataset will be included, such as the jurisdiction covered by each document.

We will provide a variety of supporting outputs:

- **Visualizations** of relationships between governance documents and taxonomy frameworks.
- A **preprint** of key findings, gaps, and opportunities for impact in the AI governance landscape.
- **Accessible content (e.g., blogs, video, social media posts)** to share findings with the broader community interested in AI risks and governance.

Acknowledgements

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

- **Graham Ryan**, Jones Walker LLP
- **Himanshu Joshi**, Vector Institute for Artificial Intelligence
- **Emre Yavuz**, Cambridge Boston Alignment Initiative
- **Sophia Lloyd George**, Brown University
- **Echo Huang**, Minerva University
- **Clelia Lacarriere**, MIT
- **Lenz Dagohoy**, Ateneo de Manila University
- **Henry Papadatos**, SaferAI
- **Aidan Homewood**, Centre for the Governance of AI

Appendix 1: Taxonomies Used

We began by extracting and comparing governance frameworks from five foundational sources, including:

- The World Bank Global Trends in AI Governance
- OECD Framework for the Classification of AI Systems
- NIST AI Risk Management Framework
- AGORA database and its label mechanisms
- MIT AI Risk Repository and Mitigation Taxonomies

We manually extracted and compared all relevant governance taxonomies, modified some to tailor them to our research context better, and gathered internal feedback within the author team.

We then tested the updated taxonomies by labeling additional governance documents both manually and with AI tools, and achieved improvements. Eventually, we settled on the following for use in our pilot.

- **AI Risk Domain**: Specific risks governed by laws, policies, or standards (per the MIT AI Risk Taxonomy).
- **AI Actors**: Entities proposing, enforcing, monitoring, or regulating.
- **AI System Lifecycle**: Stages of AI development and use .
- **Legislative Status**: Binding nature of governance (hard law, soft law, or other).
- **Sectors Governed**: Industry sectors targeted.

Appendix 2: Six Documents Reviewed

Anthropic Responsible Scaling Policy (Anthropic, 2024) (Agora ID: 768)

Representative of frontier AI companies' self-governance strategies. Responsible Scaling Policy (RSP) sets staged technical and organizational safety protocols—anchored in AI Safety Levels—to ensure the company will not train or deploy models capable of catastrophic harm without specified safeguards in place.

Anthropic first released the RSP in Sep 2023, and iterated it. The listed version was effective on October 15, 2024 (The latest version took effect on May 14, 2025).

TAKE IT DOWN Act (U.S. Senate, 2024) (Agora ID: 1293)

A Senate bill aimed at combating the online distribution of non-consensual intimate imagery—including computer-generated deepfakes—by establishing takedown duties for covered platforms and criminal penalties for offenders.

Executive Order on Removing Barriers to U.S. Leadership (White House, 2025) (Agora ID: 1780)

A presidential executive order aimed at bolstering U.S. competitiveness in AI and other emerging technologies by streamlining regulatory barriers, investing in research, strengthening domestic talent pipelines, and promoting international cooperation. It tasks federal officials with developing a comprehensive AI Action Plan for global AI competitiveness, emphasizing innovation free from ideological bias. The order revoked prior AI directives (notably EO 14110) and directed immediate review and amendment of any policies inconsistent with the revoked EO 14110. Signed January 23, 2025.

Google DeepMind Frontier Safety Framework, Version 2.0 (Google DeepMind, 2025) (Agora ID: 2040)

DeepMind's Frontier Safety Framework (FSF) v2.0 establishes Critical Capability Levels (CCLs) to assess AI risks related to catastrophic misuse and deceptive alignment. The framework requires evaluations for proximity to each CCL and mandates the application of mitigations before further scaling. It sets out detection and monitoring strategies for deceptive alignment, and defines deployment procedures to manage misuse risks. FSF 2.0 refines safety thresholds, evaluation protocols, and governance processes, representing a staged self-governance regime for frontier AI development.

California SB 53 — CalCompute and Whistleblower Protections (California Legislature, 2025) (Agora ID: 2041)

A California bill establishing governance measures for frontier AI development. The February 27, 2025 version of the bill creates a CalCompute Consortium under the Government Operations Agency to design a framework for a state-backed public cloud computing resource, advancing AI in a responsible and equitable manner. The bill also provides whistleblower protections, prohibiting large AI developers from retaliating against employees, contractors, or affiliates who disclose critical AI risks. It requires developers to establish internal, anonymous reporting channels and to notify employees of their rights, ensuring legal protections against retaliation. The bill was updated and finally passed on September 29, 2025.

San José AI Policy 1.7.12 (City of San José, 2025) (Agora ID: 2047)
The City's official policy governing municipal use of AI, setting principles including effectiveness, transparency, equity, accountability, human-centered design, privacy, security and safety, and workforce empowerment. It designates responsible bodies, defines approved and prohibited uses, and establishes sunset procedures for municipal AI systems.

Appendix 3: Dissemination and Intellectual Property Approach

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