ACLU RESEARCH REPORT

Using AI to Make Sense of AI Policy

Computational Methods to Analyze and Improve Legislative Proposals Related to AI

ACLU

BROWN Data Science Institute

CENTER FOR TECHNOLOGICAL RESPONSIBILITY, RE-IMAGINATION, AND REDESIGN

A collaboration between the Center for Tech Responsibility (CNTR) at Brown University and the American Civil Liberties Union (ACLU)

Authors: Tuan Pham, Brown University (now at Virginia Tech); Evani Radiya-Dixit, ACLU; Marissa Gerchick, ACLU; Brooke Madubuonwu, ACLU; Suresh Venkatasubramanian, Brown University

Paid for by American Civil Liberties Union, Inc., 125 Broad Street, New York, New York 10004, not authorized by any candidate or candidate's committee.

ACLU

BROWN
Data Science Institute

CENTER FOR TECHNOLOGICAL RESPONSIBILITY,
RE-IMAGINATION, AND REDESIGN

Contents

Executive Summary

In recent years, policymakers have rapidly increased their focus on regulating artificial intelligence (AI) and automated decision systems (ADS) through legislation and other policy approaches. These regulatory and legislative approaches, however, are not advancing in isolation, but often borrow and refine language from one another, creating relationships among proposals that would otherwise appear to be moving in parallel.

At the Center for Tech Responsibility (CNTR) at Brown University and the American Civil Liberties Union (ACLU), we believe carefully applying computational methods to analyze emerging AI and ADS legislation can help such entities understand this legislative landscape. In this report, we apply computational methods to 1,804 state and federal bills related to AI and ADS, introduced in state legislatures and the U.S. Congress between 2023 and mid-April 2025, both from an inter-bill and an intra-bill perspective as described below.

Key Takeaways

Inter-bill analysis: Applying computational tools across bills, we demonstrate how comparing multiple bills helps us quickly track trends over time and across states, as well as easily visualize similarities across bills to trace the overall reach of model legislation.

Intra-bill analysis: Applying computational tools within a given bill, we show how a bill's definitions can be modeled as a directed graph and how cycle detection and degree analysis can reveal potential ambiguities and highlight important definitions. We discuss how these methods can help policy staff and policymakers quickly identify how to strengthen a bill's clarity.

Technical recommendations: We provide recommendations, including that: (a) advocates and policymakers use computational tools to understand and strengthen legislation, such as by analyzing cycles to prevent loopholes; (b) researchers work with legislators and policy staff to create standardized formats for legislative texts to enable computational analysis; and (c) researchers and advocates incorporate a multilingual perspective when analyzing AI legislation introduced in regions under U.S. jurisdiction.

Analyzing AI Policy Across Bills

First, we take an inter-bill perspective and explore applying computational tools to analyze multiple bills at the same time.

Themes of AI legislation: We apply topic modeling to identify the themes of AI and ADS legislation across the United States, demonstrating how this method can give policy analysts, researchers, and technology developers a high-level view of the legislative landscape. We identify popular legislative themes in bills at the state level, many of which focus on the regulation of generative AI and creation of task forces related to AI. Policy staff can use such thematic analyses to identify policy trends and understand how these themes vary or remain consistent across jurisdictions.

Policy diffusion of AI legislation: Legislative bills often copy language from other bills or model bills (with small but sometimes significant tweaks). For example, scholars have highlighted how the "California effect" will likely play out in this context, and that AI regulation in California will have a ripple effect on other states. To better understand such policy diffusion, we apply text comparison methods to identify how bills may reuse language from model bills such as the Workday Model Bill and the Lawyers' Committee Model Bill (the Online Civil Rights Act from the Lawyers' Committee for Civil Rights Under Law). We demonstrate how this kind of text comparison can be used to identify specific bills that are highly similar to model bills amongst large samples of bills, allowing for targeted analysis by policy staff. For instance, we find that the AI Civil Rights Act of 2024 (congressional bill S.5152) is very similar to the Lawyers' Committee Model Bill. Overall, we observe that a handful of proposed bills each year from 2023 to 2025 are similar to the Lawyers' Committee Model Bill and that there has been a marked decrease in bills similar to the Workday Model Bill in 2025 compared to 2024. Policy staff can continue to track how these trends in policy diffusion unfold in upcoming legislative sessions.

Analyzing AI Policy Within a Given Bill

We also take an intra-bill perspective and explore applying computational methods using only the content of a given bill.

Applying graph theory to legislative definitions: Bills often define key terms that are used throughout them, and these definitions significantly shape the bill's scope and impact. In AI-related legislation, these definitions — of AI, AI's scope of use, and the entities held accountable — have often been a subject of contention, as they establish precedents and boundaries for AI governance. A single bill may contain several dozen definitions, many of which may reference each other, and computational methods may make understanding those interlocking definitions easier. We demonstrate how methods from graph theory can be applied to model definitions as a directed graph, which can help policy staff analyze and strengthen bill definitions.

Analyzing cyclical definitions: By visualizing definitions as a graph, we can apply graph theory methods to identify cyclical references in definitions. We provide an example of a cycle between three definitions in the AI Civil Rights Act of 2024, and discuss how sometimes cycles may be indicators of definitions that need only minor clarification, and other times may indicate significant ambiguity and possible loopholes in a bill. We argue that identifying and potentially addressing such cycles in definitions can improve a bill's clarity and reduce the likelihood of loopholes in its application.

Quantifying reliance in definitions: We also quantify the extent to which definitions rely on each other within a given bill, and demonstrate how these degrees of reliance can serve as indicators of a definition's importance in a bill, helping policy staff focus their attention and resources. For instance, a term that heavily relies on other terms in a bill may be structurally important for a bill, such as the term "sensitive covered data" in the American Privacy Rights Act. Because this term serves as an umbrella that ties together many other terms, policy staff may want to ensure its definition is clear and robust. We provide resources for interpreting a term's reliance depending on the context and nature of the bill, and policy staff can use these tools to identify which definitions to focus their attention on.

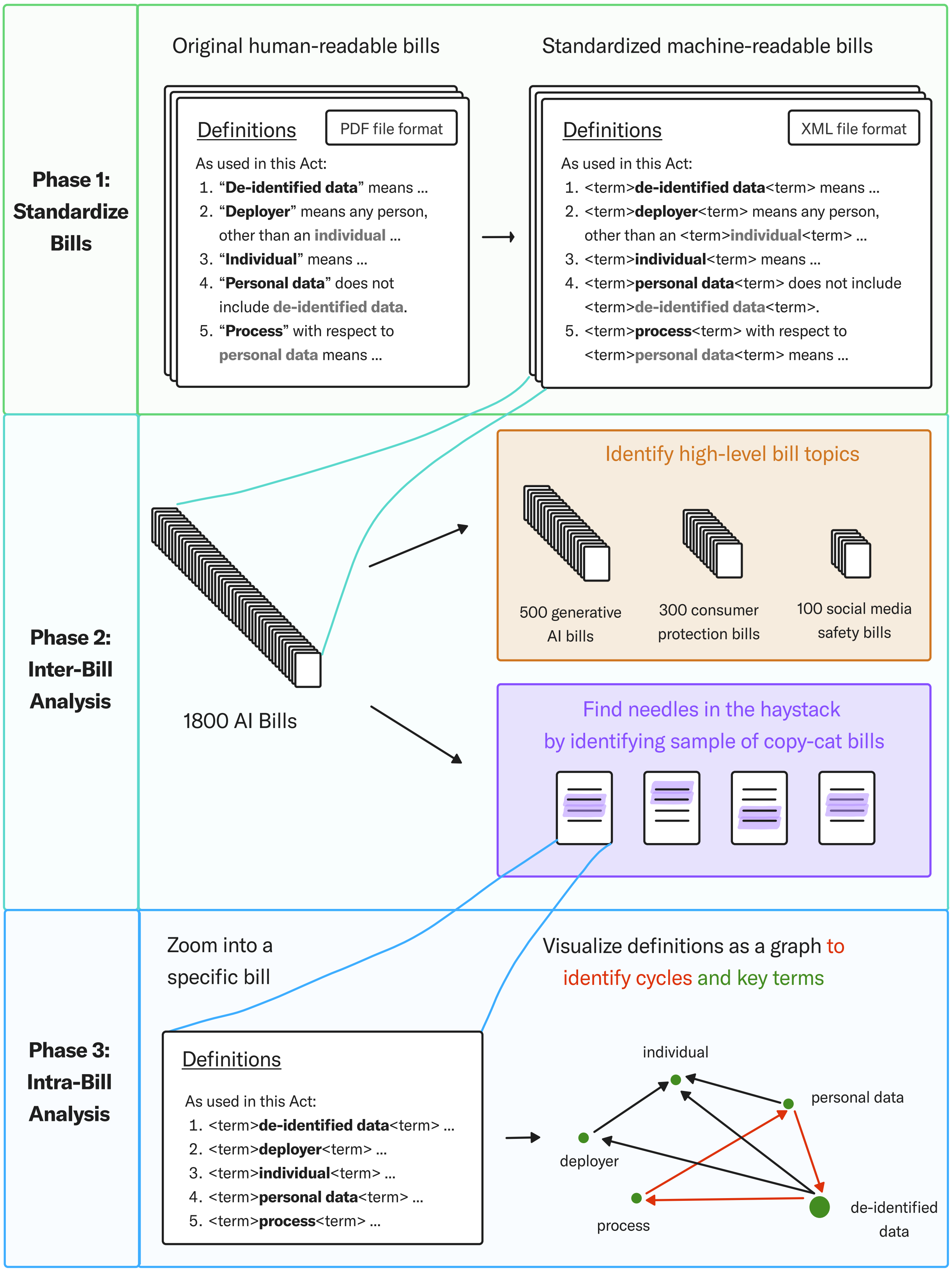
Recommendations

We provide two key recommendations to address the **technical** challenges that emerge when conducting computational AI policy analysis.

First, we recommend researchers and policy staff work together to **create standardized formats and structures for legislative texts** across jurisdictions. Establishing consistent file formats, structures of definitions and sections, annotation conventions, and references would facilitate computational analysis of legislative data and make it easier for policy staff to track changes over time.

Second, we encourage researchers and advocates to **incorporate a multilingual perspective when analyzing AI legislation** introduced in regions under U.S. jurisdiction. English-only analyses can overlook important policy developments, such as Puerto Rico's bills written in Spanish and Hawai'i's legislation sometimes written in Hawaiian. Leveraging language technologies tailored to specific languages and legal contexts, while engaging with native speakers and regional AI policy experts, would provide insights into the diverse approaches to AI policy.

User Journey for Computational AI Policy Analysis



We provide a visualization of how users can leverage computational methods for AI policy analysis. Users start with a large sample of 1,800 state and federal bills related to AI and ADS (See Section 2 for detail on how this sample of bills was constructed).

In **Phase 1**, users standardize the format and structure of bill texts for them to be machine-readable. For example, users can convert the original human-readable bills that are in a PDF file format into an XML file format, where the terms referenced in the bill's definitions are explicitly tagged (e.g., the definition for "deployer" references the term "individual"). This standardization is applied to all 1,800 bills to enable computational analysis in the following phases.

In **Phase 2**, users leverage inter-bill analysis to understand and narrow down this large set of bills. In particular, users can (a) group the 1,800 bills into high level topics, such as "generative AI," and identify the number of bills in each topic, and (b) identify bills with very similar language, also known as "copy-cat bills."

After identifying a specific bill, a needle in the haystack of 1800s bills, users can leverage intra-bill analysis in **Phase 3** to more closely examine the bill's contents. Here, users can create visuals for specific sections of the given bill. For example, users can visualize the bill's "Definition" section as a graph to identify cyclic definitions and key terms to target when strengthening the bill.

These phases of analysis build on each other and enable users to analyze and improve bills related to AI and ADS.

1. Introduction

In recent years, policymakers around the world have been increasingly interested in regulating artificial intelligence (AI) and automated decision systems (ADS). In the United States, this interest in regulation has materialized through legislation introduced at the local, state, and federal levels; federal executive orders; agency guidance and frameworks; and a variety of other policy vehicles. OpenAI's release of the consumer-facing, generative AI chatbot ChatGPT in late 2022 and the subsequent proliferation of companies building and applying large language models (LLMs) and other generative AI systems has further intensified regulatory efforts.

On the legislative front in the United States, regulators at the state and federal level have proposed many diverse approaches to regulating AI and ADS. At the state level, bills covered topics such as algorithmic discrimination, deepfakes, data privacy, algorithmic accountability, and data transparency. Amidst these dynamics, multiple policy trackers have surfaced to monitor this landscape.

At CNTR and the ACLU, we believe carefully applying computational methods to analyze emerging AI and ADS legislation can help such entities and organizations understand this landscape. In this report, we apply computational methods to 1,804 state and federal bills, introduced in state legislatures and the U.S. Congress between 2023 and mid-April 2025 (See Section 2 for more detail on how this sample of bills was constructed), from two different perspectives:

1. ***Analyzing the AI policy landscape from an inter-bill perspective:*** First, we draw insights from looking at multiple bills at the same time, taking an inter-bill perspective to analyze legislative content. We demonstrate the utility of topic modeling for thematic analysis (**Section 2.2**), and highlight how comparing bills can provide insights into policy diffusion (**Section 2.3**).
2. ***Analyzing AI policy from an intra-bill perspective:*** Second, we explore how we can apply computational methods using only the content of a given bill. We demonstrate how definitions in a bill can be modeled as a directed graph (**Section 3.1**), and how graph theory methods and metrics like cycle detection (**Section 3.2**) and degree analysis (**Section 3.3**) can serve as valuable resources for policy staff, providing granular insights on the clarity and precision of particular legislative proposals.

We conclude with an assessment of the potential to use these and other data analysis techniques as tools to support and increase access to AI policy discussions. We also provide two key recommendations for: (1) legislators and policy staff to work together to create standardized formats for legislative texts that make computational analysis of this data easier (**Section 4**) and (2) researchers and advocates to incorporate a multilingual perspective when analyzing AI legislation introduced in regions under U.S. jurisdiction (**Section 5**). Finally, we consider the methods and examples in this report to be early case studies in a broader legislative analysis effort, and we conclude with a discussion of our plans to expand on this work.

Throughout this work, we take an approach grounded in carefully selecting the right tool (computational or otherwise) for the job at hand — as opposed to more haphazard approaches, such as indiscriminately applying the latest generative model for a given task — and balancing the benefits of using such computational tools against the possible risks.

2. Analyzing the AI Policy Landscape From an Inter-Bill Perspective

When analyzing AI policy, we can take an *inter-bill* perspective to help policy analysts, journalists, researchers, and other policy staff gain a macroscopic view of the current legislative landscape across the United States. This macroscopic view is more important than ever given the rapid increase in the number of AI and ADS bills being introduced and advanced around the country, including both at the state and federal level. Inter-bill analyses can help answer questions such as:

- Which states are most active in producing and enacting AI and ADS legislation?
- Within the universe of possible AI and ADS legislation, how are particular topics or themes distributed across states and at the federal level?
- How often is language shared across AI and ADS legislation, and what does this language sharing look like?

We highlight the possibilities of inter-bill analysis in the next several sections, demonstrating the value of topic modeling for thematic analysis of bill text and text comparisons for analyzing language similarity in legislation.

2.1. Defining the scope of our inter-bill analysis

For these analyses, we used data resources from the Open States project operated by Plural Policy — a bill tracking platform with open access to federal and state legislative data — to obtain bill texts and metadata (e.g., the bill's sponsors, key events in the legislative history of a bill, dates associated with those events, etc.) of AI and ADS bills at both the state and federal levels.

To define the scope of relevant AI and ADS legislation, we analyzed bills that:

1. were introduced between January 2023 and mid-April 2025[1] (for each bill, we analyze only the most current version); and
2. contain a set of selected keywords related to AI and ADS, (including "artificial intelligence," "automated decision systems," "deep fake," "facial recognition," and "large language model") or close variations of these keywords;[2] and
3. meet one of the following criteria: (a) contain "artificial intelligence," (b) contain "automatic[3] decision," or (c) meet a certain threshold of the number of matched keywords (from 2) or number of keywords relative to the length of the bill[4]. This criteria is designed to filter out bills that mention these keywords only tangentially or extraneously. For instance, Arizona HB 2889 is a bill related to pornography that mentions "facial recognition" once, but reading the bill makes it clear that facial recognition technology is not a substantive focus of the bill, and the bill has no other keywords of interest (from 2).

Amongst those bills that meet the aforementioned criteria, we excluded:

1. budget- or fiscal-related bills[5], as they often cover a wide range of topics and are qualitatively different from most of the other bills we analyze here. There are currently around 89 of these bills (which satisfy the above conditions but contain budget-related keywords in the title); and

2. bills introduced in the Senate or House of Representatives of Puerto Rico, which are generally in Spanish and for which government-approved English translations do not appear to be available. These bills show up in our data since Puerto Rico falls under U.S. jurisdiction as a U.S. colony. Although our tooling is currently not equipped to handle non-English texts, see **Section 5** for our call for multilingual AI policy analysis.[6]

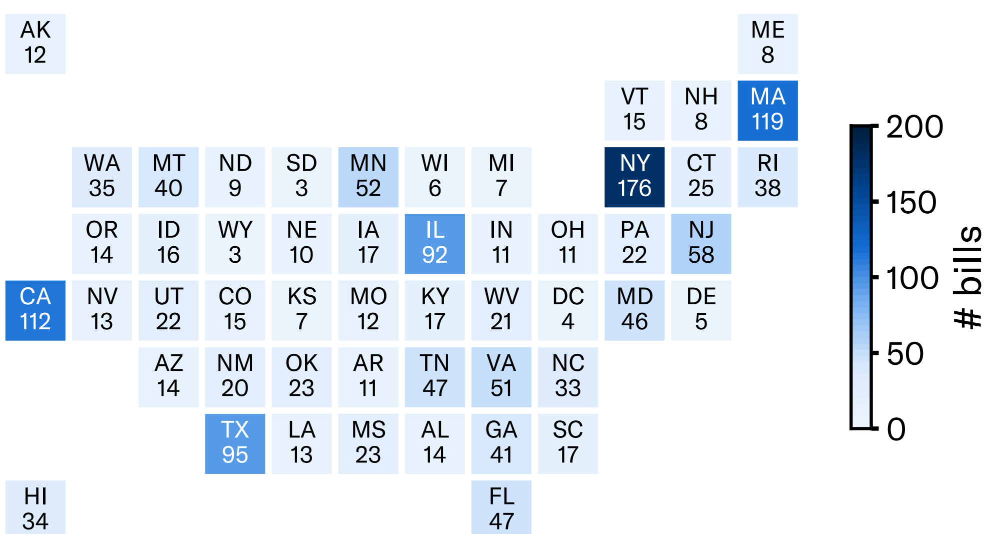
While some efforts to track and quantify the prevalence of these kinds of bills to date have used simpler criteria to identify relevant legislation (for example, some reports we identified defined "AI-related legislation" as any bill mentioning "artificial intelligence")[7], we believe this approach can help more accurately identify legislation related to AI and ADS that may not explicitly use these terms — such as bills about types of AI like facial recognition technology — while excluding legislation that touches on these issues only tangentially. Importantly, as discussed in the next section, the bills in this sample have varying degrees of substantive requirements and applicability; for example, roughly 30% focus on establishing task forces, and many bills focus only on uses of AI in particular sectors, such as healthcare.

In total, between 2023 and mid-April 2025, there were 1,804 bills introduced that meet these criteria. Out of these bills, 240 are federal bills, and 1,564 are state bills (including bills from Washington, D.C.) (see **Figure 1** for more detail on the geographic distribution of introduced bills in our sample). The three states with the most bills introduced are New York (176), Massachusetts (119), and California (112).

For our analysis, we performed some pre-processing, including text cleaning (e.g. to remove line numbers and PDF artifacts, and handle other formatting nuances to the best of our ability). A substantial portion of the bills in our sample are quite lengthy (about 14% of the bills in our sample, 245 of the 1,804, are longer than 5,000 words), and oftentimes only a few sections of these lengthy bills mention the relevant keywords.[8] To appropriately scope our analysis, we trimmed bills longer than 5,000 words, so only portions of the bills around our anchor keywords remained in our sample for further analysis.[9] Of the 245/1,804 bills we trimmed using this procedure, on average, 23% of the word count of the full text remained in our sample after trimming. Taking all the 1,804 bills together, the processed content is on average around 89% of the word count of the original texts.

FIGURE 1

Bills pertaining to AI or ADS (January 2023 - April 2025)
Total 1804 bills (240 Congressional bills, 1564 State bills)



Number of bills pertaining to AI and Automated Decision Systems (ADS) introduced per jurisdiction (including Washington, D.C.) between January 2023 and April 2025.

2.2. Performing thematic analysis using topic modeling

The first approach we used to analyze the legislative landscape of our sample was to perform thematic analysis on the content of these 1,804 bills to categorize their focuses and approaches. While states sometimes tag proposed legislation with certain subject matters or topics, such information can be sparse, inconsistent across jurisdictions, and of varying levels of specificity/usefulness. For example, some bills on state legislatures' websites, such as Maryland HB 1331 (2025), are tagged with a wide-ranging set of topics on their state legislature websites (some as specific as "Artificial Intelligence," and others as general as "Time"), while others, like Vermont H 341 (2025-2026), are not tagged with topics at all.

Given the need to apply a consistent thematic categorization method to these bills, we used a topic modeling approach inspired by the 2023 Global AI Infrastructures Report by George Mason University and the Stimson Center. Specifically, we performed topic modeling using *ensemble Latent Dirichlet allocation (eLDA)* on legislative texts. At a high level, by aggregating text from all these bills, this method creates abstract semantic clusters (i.e. *topics*) in a robust manner, and then assigns each bill a probability of belonging to each of the extracted topics, providing a method of describing each bill as a mixture of different topics. Using this method, we allowed for a bill to be assigned to multiple topics.

Applying this method, we organized the 1,804 bills in our sample into 12 topics. The word representation of each topic is shown in **Figure A1** in the Appendix. For each bill, we examined whether it was assigned to at least one topic with at least 25% probability: 1,759 out of 1,804 (97.5%) bills met this criteria, and we excluded the remaining 45 bills from further thematic analysis.[10] For the 1,759 bills, we then manually grouped the 12 topics they were assigned to into seven clusters since several topics had overlapping themes (see **Figure A1** and **Table A1**). **Table 1** below describes these seven legislative clusters.

TABLE 1

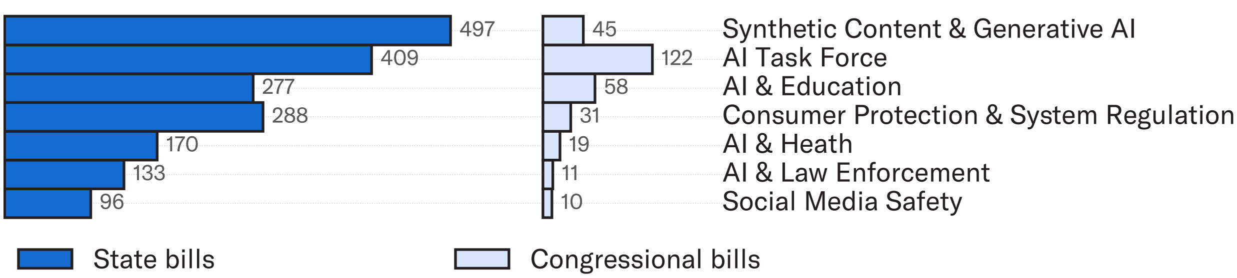
Summary of Legislative Clusters Used to Categorize Bills

Legislative Clusters & Example bill titles	Description
Consumer Protection & System Regulation Examples: • New York S 2277 (2023-2024): Enacts the "digital fairness act" • Maryland HB 1331 (2025): Consumer Protection - Artificial Intelligence • Vermont H 341 (2025-2026): An act relating to creating oversight and safety standards for developers and deployers of inherently dangerous artificial intelligence systems	Legislation imposing consumer-protection-focused requirements (including impact assessments, audits, and notice) on developers and deployers of AI and algorithmic systems used in high-risk areas.
AI Task Force Examples: • US S 1356 (118): ASSESS AI Act • Nevada SB 165 (82): Revises provisions relating to businesses engaged in the development of emerging technologies. (BDR 18-878) • Connecticut HB 5047 (2025): AN ACT CREATING A TASK FORCE TO STUDY ARTIFICIAL INTELLIGENCE AND THE STATE WORKFORCE.	Legislation requiring the creation or convening of task forces to evaluate general or specific (e.g., in a particular industry or geographic area) impacts of AI.
Synthetic Content & Generative AI Examples: • Louisiana SB 97 (2024): POLITICAL CAMPAIGNS: Regulates the use of deep fakes and artificial intelligence technology in political advertising. • Ohio SB 163 (136): Regards AI images, simulated child porn, replica identity fraud • New York A 6540 (2025-2026): Requires generative artificial intelligence providers to include provenance data on certain content made available by the provider	Legislation related to the creation and dissemination of synthetic content online created using AI, including deepfakes. This legislation often focuses on the risks of deepfakes in elections or the creation and dissemination of explicit imagery, sometimes requiring watermarking or other provenance measures for AI-generated content.

AI & Law Enforcement Examples: • New York A 10625 (2023-2024): Relation to the regulation of the use of artificial intelligence and facial recognition technology in criminal investigations • Montana LC 1312 (2025): Prohibit facial recognition technology at traffic lights • Illinois HB 3882 (103rd): STANDARD ID-DOCUMENTATION	Legislation regulating the development and use of facial recognition, facial analysis, or other biometric technologies, often focusing on uses by government agencies in identification, policing, and immigration contexts.
AI & Health Examples: • Virginia SB 392 (2024): Hospitals; emergency departments to have at least one licensed physician on duty at all times. • Rhode Island HB 5172 (2025): AN ACT RELATING TO INSURANCE – THE TRANSPARENCY AND ACCOUNTABILITY IN ARTIFICIAL INTELLIGENCE USE BY HEALTH INSURERS TO MANAGE COVERAGE AND CLAIMS ACT • Texas SB 1411 (89R): Relating to the use of artificial intelligence-based algorithms by health benefit plan issuers, utilization review agents, health care providers, and physicians.	Legislation with varying approaches to regulating the use of AI in health contexts, including requiring health care providers to allow patients to use AI assistants in health care settings and regulating how insurers and health care entities can use AI for decision-making.
AI & Education Examples • Missouri HB 2612 (2024): Establishes an educational technology impact advisory council to review the use of technology in schools • Illinois SB 1677 (104th): SCH CD-TEACHER EVALUATION PLAN • Tennessee HB 933 (114): AN ACT to amend Tennessee Code Annotated, Title 49, relative to school safety.	Legislation related to the use or evaluation of AI in education contexts, including uses of AI in the classroom, AI education for students, and the use of AI for non-educational uses in schools, such as for attempted weapons detection.
Social Media Safety Examples: • North Carolina SB 514 (2025): Social Media Control in IT Act. • Illinois HB 3943 (103rd): SOCIAL MEDIA MODERATION • Oklahoma SB 885 (2025): Social media; creating the Safe Screens for Kids Act. Effective date.	Legislation related to social media, often implementing restrictions designed to regulate minors' use of social media.

FIGURE 2

Number of Bills Per Legislative Cluster



Comparison of state vs. congressional bills across legislative clusters. For each legislative cluster, the dark blue bar plot (left) shows the number of bills at the state level, including bills from Washington, D.C., and the light blue bar plot (right) shows the number of bills at the congressional level.

Legislation by the numbers*:* The two most popular legislative clusters in our sample are (1) *Synthetic Content & Generative AI* and (2) *AI Task Force*, each with more than 500 state and congressional bills (**Figure 2** and **Table 2**). The next most popular clusters are (3) *AI & Education* and (4) *Consumer Protection & System Regulation*, each with more than 300 bills. The remaining clusters are generally more domain-specific, each with 100-200 bills per cluster: (5) *AI & Health*, (6) *AI & Law Enforcement* and (7) *Social Media Safety.*

TABLE 2

Summary of Geographic Diversity of Legislative Clusters

Legislative Clusters	Number of States with Bills	Total Number of Bills (State and Congressional)	% of Bills Introduced at State Level	Top 3 States for Bills* (including Washington, D.C.) * Also see Figure A3. for maps of clusters
AI Task Force	47	531	77%	Massachusetts, New York, New Jersey
AI & Education	41	335	83%	Texas, Massachusetts, Florida
AI & Health	33	189	90%	Texas, Virginia, Massachusetts
Consumer Protection & System Regulation	37	319	90%	New York, California, Massachusetts
Social Media Safety	24	106	91%	New York, Massachusetts, California
Synthetic Content & Generative AI	51	542	92%	New York, California, Texas
AI & Law Enforcement	32	144	92%	Illinois, Massachusetts, New York

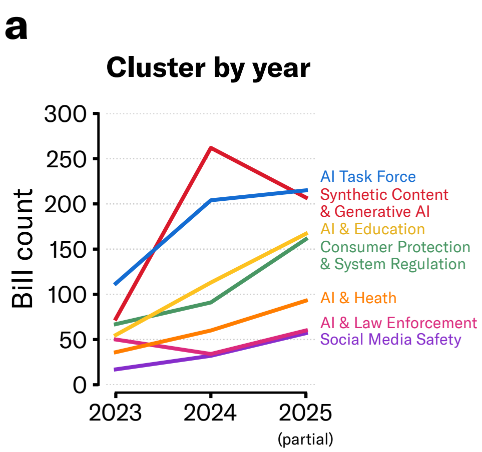
Geographic diversity of legislation: Of the two most popular clusters, (1) *Synthetic Content & Generative AI* and (2) *AI Task Force*, *Synthetic Content & Generative AI* bills were more likely to be introduced at the state level than *AI Task Force* bills (**Figure 2**, **Table 2**). Specifically, about 92% of bills related to *Synthetic Content & Generative AI* were introduced at the state level in all 51 jurisdictions (including Washington, D.C.), with the state of New York alone introducing more bills in this cluster than Congress (**Table 2, Figure A3**). In contrast, about 77% of bills related to *AI Task Force* were introduced at the state level, and more than 100 bills were introduced in Congress on this matter — more than in any single state. We also observed a similar trend with *AI & Education* — there were more bills introduced in Congress on this matter than in any individual state.

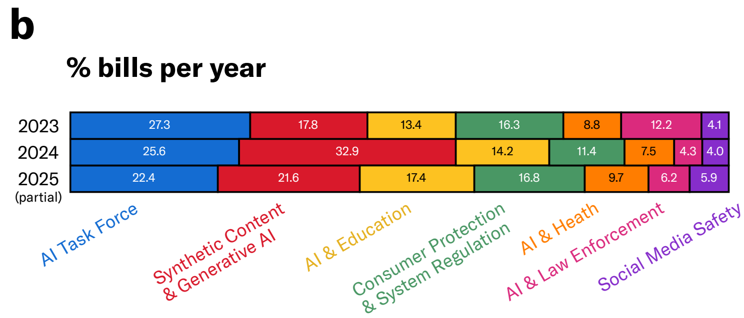
For the remaining clusters, states are clearly leading the way on introducing bills in these areas, with more than 90% of bills introduced at the state level per cluster. Additionally, in **Table 2**, we see clear trends in the states that most frequently introduce legislation in these clusters. For instance, New York, Massachusetts, California, and Texas are consistently leading the states in introduction of these bills, with Illinois, Florida, Virginia, and New Jersey emerging as hotspots for AI- and ADS-related legislation as well.

Legislation across time: When analyzing bills from the beginning of 2023 to April 2025, we observed that the two most popular clusters, *AI Task Force* and *Synthetic Content & Generative AI* dominated legislative interests across the time period we analyzed, both by sheer number of bills (**Figure 3a**) and by percentage of bills per year (**Figure 3b**). In the first few months of 2025, the number of *Synthetic Content & Generative AI* bills and the number of *AI Task Force* bills introduced were relatively similar. The other legislative clusters also saw increases in the number of bills introduced across the years, except for AI *& Law Enforcement,* which had a slight decrease in 2024 and then a rebound in the first quarter of 2025.

FIGURE 3

Legislative Topic Clusters by Year





Topic clusters of all state and congressional bills by year from 2023 to April 2025. (**a**) Number of bills for each cluster by year. (**b**) Percentage of bills for each cluster by year (with 2025 being a partial year).

In this section, we illustrate how topic modeling can be used to identify the themes of AI and ADS legislation across the United States, giving policy analysts, researchers, and technology developers a high-level view into the legislative landscape. These results can help identify popular legislative interests at the state level, such as *Synthetic Content & Generative AI* bills and *AI Task Force bills*, and how these interests vary or remain consistent across jurisdictions.

2.3. Quantifying policy diffusion using pairwise bill comparisons

In addition to high-level thematic analysis, we can also analyze AI and ADS legislation at a more granular level by comparing bills directly with each other. Pairwise bill comparison, applied over a large range of bills, can allow policy staff to detect and understand the diffusion of specific text from one bill to another and the lineage of legislative language across different bills over time. Policy staff, when facing a deluge of AI- and ADS-related bills (potentially hundreds or thousands), can use this comparison method to quickly and robustly identify potential copycats of bills they know are of particular interest, allowing them to focus their time and energy accordingly. In essence, as we demonstrate below, this inter-bill analysis can help policy staff find needles in the haystack of AI- and ADS-related bills, which can then be supplemented with in-depth intra-bill analysis of those identified bills.

> Topic modeling can help identify the themes of AI legislation, giving policy analysts, researchers, and tech developers a high-level view into the legislative landscape.

This inter-bill analysis helps policy staff find needles in the haystack of AI-related bills, which can then be examined more deeply through intra-bill analysis.

In this section, we compare AI and ADS legislation with the following three model bills ("model bill" is a term used to refer to a general template for legislation that can be adapted by legislative bodies):

1. **Workday Model Bill:** This model bill, obtained by Recorded Future News, was reportedly developed and promoted by the large human resources company Workday. As highlighted in news coverage from early 2024, text that matches large portions of this model bill has shown up in several bills introduced in state legislatures.
2. **Lawyers' Committee Model Bill:** This refers to the Online Civil Rights Act from the Lawyers' Committee for Civil Rights Under Law, released in December 2023.
3. **ALEC Model Bill:** This refers to the Model State Artificial Intelligence Act from the American Legislative Exchange Council (ALEC), finalized in August 2024.

We note that our use of the term "model bill" — which is a frequently used term for templatized legislation — does not imply that a bill is good or bad, only that it is a text after which other bills may "model" their language.

To examine potential policy diffusion between bills, we first segmented each bill document into sentences (Frohmann et al 2024, arXiv). We compared sentences of at least 10 words across documents using the fuzzy partial ratio metric (rapidfuzz). This metric takes into account the number of insertions and deletions that turn one sentence into another, allowing for a tractable comparison between sentences given the fuzzy nature of text data.

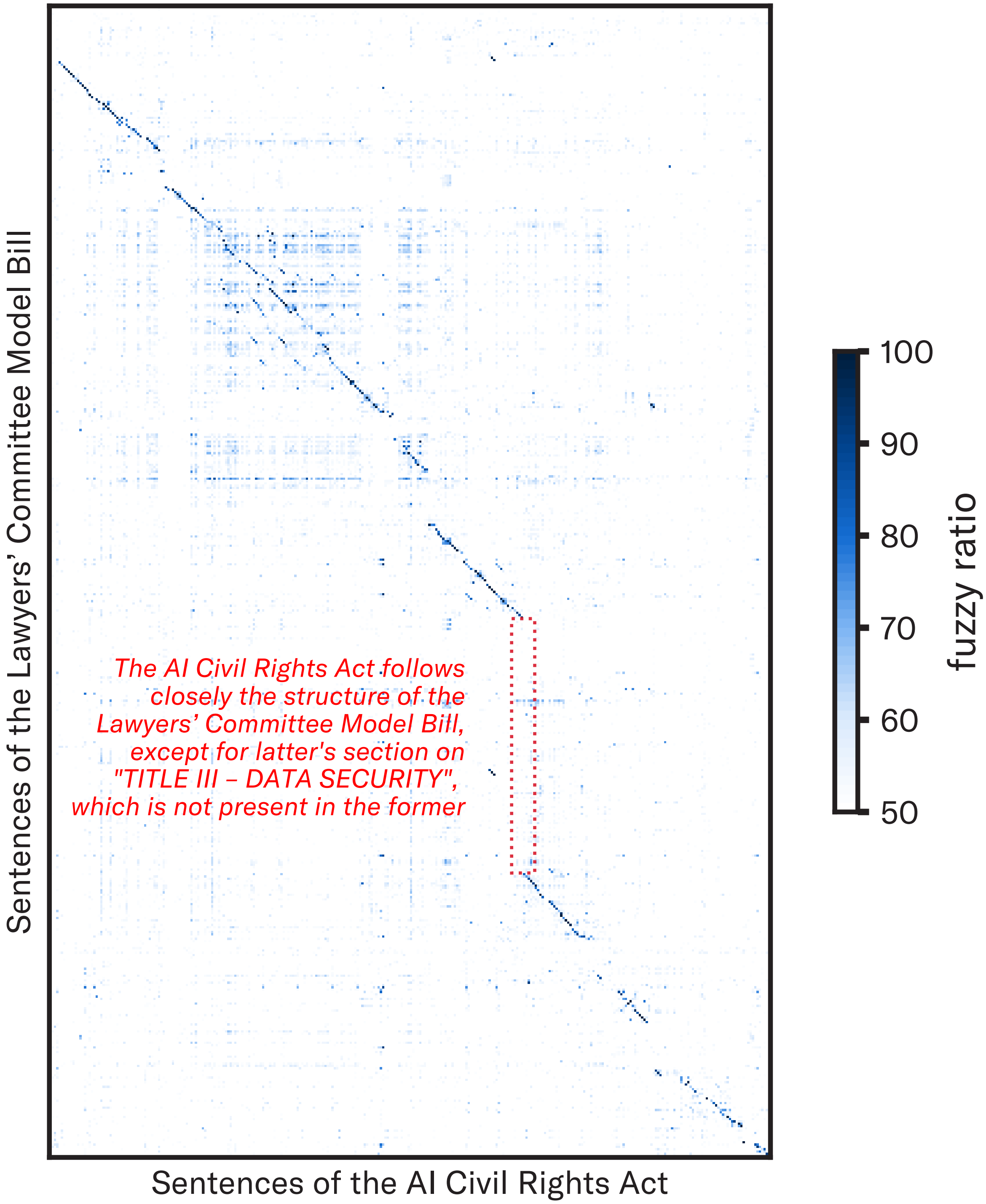
We applied this text comparison to the Lawyers' Committee Model Bill and the congressional bill S.5152, Artificial Intelligence (AI) Civil Rights Act. The congressional bill was introduced at the end of September 2024 and endorsed by the Lawyers' Committee and the ACLU. The pairwise sentence comparison between these two bills, shown in the **Figure 4** heatmap, not only confirms that they are similar structurally in an almost one-to-one linear[11] manner, but also shows that the section "Title III - Data Security" from the Lawyers' Committee Model Bill did not carry over to the AI Civil Rights Act of 2024 (see the red rectangle in **Figure 4** where there is a "break" between the blue streak).

Next, we performed text comparison to search for a given model bill's sentence within our legislation database, which allowed for the discovery of similar bills. Here, we compared the 1,804 bills discussed in **Section 2.1** with the three aforementioned model bills. We categorized a legislative text as "highly similar" to a given model bill if there were at least 10 sentences in the model bill with at least 80% partial fuzzy ratio when compared to the legislative text (see **Figure A4** for the distribution of highly similar model bill sentences and an explanation of this threshold). **Figure 5** shows the number of pieces of legislation that have a high similarity with model bills across years and jurisdictions.

We found that the Lawyers' Committee Model Bill shows similarity with proposed legislation consistently across years, with a slight increase in

FIGURE 4

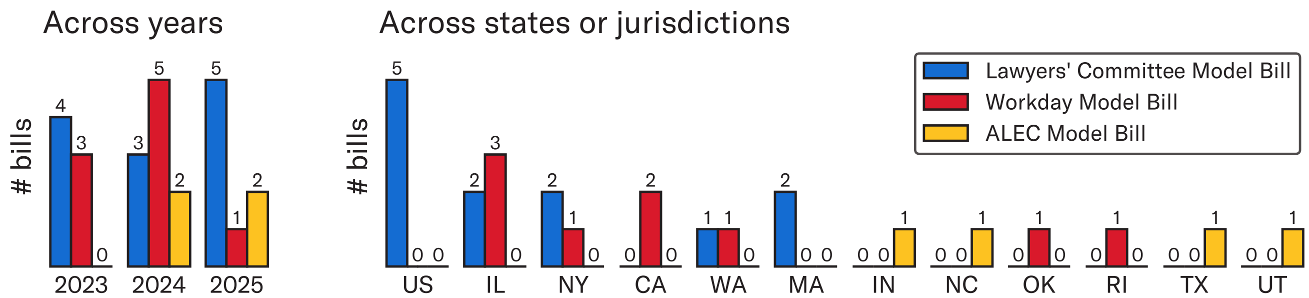
Comparisons between tech civil rights acts (only showing sentences with ≥ 10 words)



Demonstration of sentence-wise comparisons between tech-related civil rights acts. The heatmap shows how similar sentences of the legislative text from the AI Civil Rights Act of 2024 (x-axis) are to those of the Lawyers' Committee Model Bill (y-axis). Only sentences with at least 10 words were compared and shown in this heatmap. Fuzzy ratios (see text for methodology) larger than 80% typically indicate similarity. To interactively inspect these sentence pairs between these two bills (and additionally NY S-5641 and the Workday Model Bill), please view at: https://brown-cntr.github.io/standalone-htmls/ai-leg/demo-pairwise-sentence-bill-compare.html.

FIGURE 5

Counts of bills that share similarity with ≥ 10 sentences from different model bills



Number of bills that share similarity with model bills. Text from a bill is considered "highly similar" with a model bill if at least 10 sentences from the latter with at least 10 words reach the 80% partial fuzzy ratio when compared to the former (the 10-sentence threshold was chosen to consistent with all three templates, see the distributions in Figure A4). The number of bills is tabulated for each template (colors), across years (left, with 2025 being a partial year) and across jurisdictions (right; Congress = US; States, including Washington, D.C., are shown on the right of Congress).

the beginning of 2025. Many congressional bills, in addition to the AI Civil Rights Act of 2024, as well as bills from Illinois, New York, Washington, and Massachusetts display high similarity to this model bill. In contrast, the Workday Model Bill shares similarity with only state bills. A 2024 investigation from The Record found similarities between the Workday Model Bill and six state legislative bills. Our analysis largely replicated these findings, with a few exceptions: We found that the Connecticut bill SB-2 does not meet our similarity threshold requirements with the Workday Model Bill, and we identified an additional bill, HB 3835, from Oklahoma, that does. The number of bills sharing high similarity with the Workday Model Bill peaked in 2024[12] and decreased in the first few months of 2025. Lastly, there are only four bills in our data that share high similarity with the ALEC Model Bill, which may be due to the fact that the ALEC Model Bill is relatively new compared to the other two model bills. All four were introduced after 2023 from states that did not introduce bills similar to the other two model bills.

In summary, we observed consistent similarity of legislation with the Lawyers' Committee Model Bill across the years analyzed, and it is possible that bills introduced across states and Congress will continue to share similar passages with this model bill. To a lesser extent and only at the state level, we also observed similarity of legislation with the ALEC Model Bill. Since this is the newest model bill out of the three discussed here, it will be interesting to monitor whether more state bills are modeled after the ALEC Model Bill in the next legislative session. On the other hand, we noted a decrease of similarity to the Workday Model Bill in 2025. It is possible that the media coverage about and advocacy opposing this model bill in 2024, at least partially, contributed to this decline. These insights could aid policy analysts in tracking down text reuse and policy diffusion in legislation, both in the cases where certain legislative passages have been documented to be useful or contentious and in cases where analysis or coverage of legislation does not exist.

3. Analyzing AI Policy From an Intra-Bill Perspective

In addition to the macroscopic view offered by *inter-bill* analysis, we can take an *intra-bill* perspective to examine a given bill or law in detail. Intra-bill analysis can surface insights for drafters and reviewers of legislation by providing them with metrics and visualizations of complex aspects of a bill that help them focus their time and energy. In this section, we discuss how computational tools from graph theory can help policy staff make sense of legislative language.

3.1. Applying graph theory to legislative definitions

Many pieces of draft legislation include definitions of various terms used throughout the bill, often in a designated "Definitions" section. These definitions are key to determining the scope and impact of the proposal. Legislation related to AI has often been a battleground over definitions — of AI, the scope of use, and the entities held accountable — (as demonstrated by debates over California SB 1047, for instance) because these definitions establish precedents and boundaries for AI governance.

The definitions of a bill may reference one another, creating interlocking and sometimes circular definitions. Oftentimes, when reading these definitions, a policy analyst may try to mentally map out how they relate to each other, making note of which definitions reference each other — and how — while analyzing the bill. Given the complexity of tracking definitions, we explore using tools from graph theory to automatically generate a graph of the relationships between definitions.

Consider the following excerpted definitions of the terms "personal data" and "process" from the AI Civil Rights Act of 2024, introduced in the U.S. Congress in 2024:

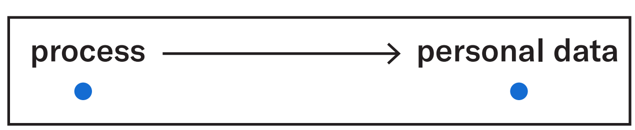
> PERSONAL DATA.— (A) IN GENERAL.— The term "personal data"— (i) means information that identifies or is linked or reasonably linkable, alone or in combination with other information, to an individual or an individual's device; and (ii) shall include derived data and unique persistent identifiers. (B) EXCLUSION.— The term "personal data" does not include de-identified data.
>
> PROCESS.— The term "process", with respect to personal data, means to conduct or direct any operation or set of operations performed on such data, including analyzing, organizing, structuring, retaining, storing, using, or otherwise handling such data.

We can think of these terms as nodes in a graph. Since the definition of "process" references the definition of "personal data," we can add a directed edge to the graph from "process" to "personal data" as shown below in **Figure 6**.

Extending this concept to the whole bill, we can create a node for each term in the bill, with an edge from one term (or node) x to another term y whenever the definition of x references y. **Figure 7** includes a graph for all definitions in the AI Civil Rights Act of 2024. Terms that have many arrows

FIGURE 6

Demonstration of modeling the relationship between two definitions with nodes and a directed edge



pointing toward them — such as "individual," "deployer," and "covered algorithm" in the AI Civil Rights Act of 2024 — have many definitions that rely on them, and thus may be more structurally important for the bill. In contrast, terms that have many arrows pointing away from them rely on many other definitions, and so may have many dependencies.

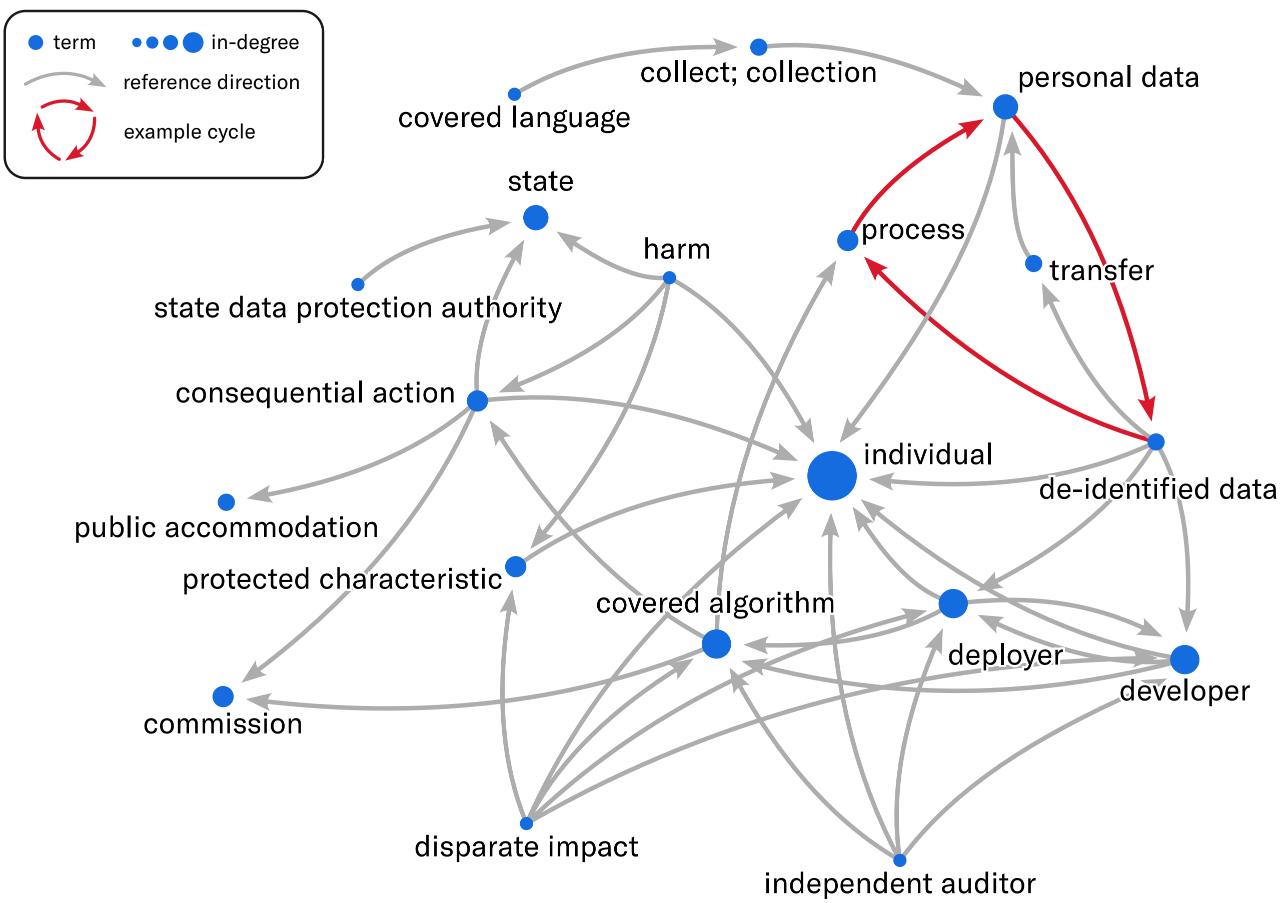
3.2. Cycle or loophole: Analyzing cyclical references in definitions

By visualizing the definitions and their relationships as a directed graph, we can apply methods from graph theory to ask questions like: Are there cyclical references in definitions? For example, the Fair Credit Reporting Act (FCRA), which provides important protections related to credit reports and background checks, defines a "consumer report" and a "consumer reporting agency" in part as follows:

> The term "consumer report" means "any written, oral, or other communication of any information by a consumer reporting agency bearing on a consumer's credit worthiness, credit standing, credit capacity, character, general reputation, personal characteristics, or mode of living..."

FIGURE 7

Directed graph of defined terms in the AI Civil Rights Act of 2024



Demonstration of modeling all terms defined in a bill, in this case, the AI Civil Rights Act of 2024, as a directed graph, with nodes sized by their in-degrees. The cycle between "personal data," "process," and "de-identified data" is colored in red. See the interactive version of the figure to explore these definitions and their relationships with each other, as well as alternative visualizations of the node sizes.

> The term “consumer reporting agency” means “any person which, for monetary fees, dues, or on a cooperative nonprofit basis, regularly engages in whole or in part in the practice of assembling or evaluating consumer credit information or other information on consumers for the purpose of furnishing consumer reports to third parties...”

According to these definitions, a report is a “consumer report” only if it is produced by a “consumer reporting agency,” and the definition of “consumer reporting agency” relies on the definition of “consumer report.” This example demonstrates how cycles in definitions can be an issue, including because they can contribute to ambiguity in the scope or applicability of a bill’s definitions. These definitions have been a source of significant debate for several reasons, including due to ambiguities created by the cyclical reference. For instance, the Consumer Financial Protection Bureau (CFPB) recognized this circularity in a recent consideration of whether the FCRA applies to new business models such as data brokerage, highlighting how data brokers have interpreted these definitions to argue they are not subject to some FCRA requirements, as they are not consumer reporting agencies selling consumer reports.

But not all cycles are equally problematic — some can be fixed with minor language clarifications or by removing extraneous phrases or clauses. For example, in the AI Civil Rights Act of 2024, the definition of “process” references “personal data,” and the definition of “personal data” references the term “de-identified data,” which itself relies on the definition of “process” — creating a cycle, as shown in red in **Figure 7**. In this instance, removing the direct reference to “personal data” in the definition of “process” can break the cycle:

> PROCESS.—The term “process”, with respect to ~~personal~~ data, means to conduct or direct any operation or set of operations performed on such data, including analyzing, organizing, structuring, retaining, storing, using, or otherwise handling such data.

Identifying and potentially addressing cycles in definitions can improve a bill’s clarity and reduce ambiguities or loopholes in its application.

Alongside this change, the bill could specify the kind of data (e.g., “personal data”) it is referring to whenever the term “process” is used in practice in the bill.

Identifying and potentially addressing cycles in definitions can improve a bill’s clarity and reduce the likelihood of ambiguities or loopholes in its application. Definitions in legislation considerably shape a policy’s ultimate impacts, and are often targeted by lobbyists seeking to influence a bill’s coverage or control its applicability. Thus, mapping definitions as a graph can serve as one useful tool, among others, for policymakers and policy staff to quickly identify and analyze cycles in definitions that may have been introduced unintentionally, and ultimately strengthen those definitions. Building on this analysis, we next discuss how additional graph theory tools and metrics can be useful for policy analysis.

TABLE 3

Terms most frequently referenced in other definitions in the American Privacy Rights Act of 2024 (APRA)

Term	Definition	Number of times cited in other definitions (up to a maximum of 59)
individual	The term "individual" means a natural person residing in the United States.	36
covered data	The term "covered data" means information that identifies or is linked or reasonably linkable, alone or in combination with other information, to an individual or a device that identifies or is linked or reasonably linkable to 1 or more individuals. *Note: This definition has various exclusions. See bill text for the full definition.*	26
collect	The term "collect" means, with respect to covered data, to buy, rent, gather, obtain, receive, access, or otherwise acquire the covered data by any means.	19
process	The term "process" means, with respect to covered data, any operation or set of operations performed on the covered data, including analyzing, organizing, structuring, using, modifying, or otherwise handling the covered data.	17
transfer	The term "transfer" means, with respect to covered data, to disclose, release, share, disseminate, make available, sell, rent, or license the covered data (orally, in writing, electronically, or by any other means) for consideration of any kind or for a commercial purpose.	17

3.3. Degrees of separation: Quantifying reliance in definitions

Alongside modeling definitions as a graph, looking at the *in-degree* (the number of edges going into a node) and *out-degree* (the number of edges coming out of a node) can provide valuable information about definitions. This kind of analysis can help policy staff identify terms that may be more or less structurally important for a bill, which can help them focus their attention and resources.

Take, for example, the American Privacy Rights Act of 2024 (APRA), which includes 60 defined terms. If we create a directed definitions graph for this bill as we did for the AI Civil Rights Act of 2024 — with nodes for each term and edges from one term to another if the first term relies on the second term — we can identify which terms have the highest *in-degree* (indicating other terms reference them frequently) and *out-degree* (indicating they reference other terms frequently) in the graph. We list the definitions with the highest in-degrees in the APRA in **Table 3** below.

If a term has a relatively high in-degree, this indicates the term is heavily relied on by others in the bill, perhaps suggesting such a term is structurally important for the bill. In the APRA, the terms "individual," "covered data," and several verbs related to operations on covered data ("collect," "process," and "transfer") are the most frequently referenced in other definitions. A low in-degree indicates a term is not frequently relied on by others in the bill and may be clear without further context. For instance, the term "parent" — which is defined as "a legal guardian" in the APRA — has an in-degree of zero.

If a term has a high out-degree (e.g., it references many other terms in the bill), this may suggest the term is complex and has many dependencies. We provide a list of the terms with the highest out-degree in the APRA in **Table 4**. This kind of metric

TABLE 4

Terms with the most references to other terms in the American Privacy Rights Act of 2024

Term	Definition (abbreviated; see bill text for full definitions)	Number of other terms referenced (up to a maximum of 58)
sensitive covered data	The term "sensitive covered data" means the following forms of covered data: a government-issued identifier; health or genetic information; financial account or card numbers; biometric information; precise geolocation information; the private communications of an individual; unencrypted or unredacted account or device log-in credentials; etc.	18
affirmative express consent	The term "affirmative express consent" means an affirmative act by an individual that—clearly communicates the authorization of the individual for an act or practice; and is provided in response to a specific request from a covered entity, or a service provider on behalf of a covered entity.	17
publicly available information	The term "publicly available information" means any information that a covered entity has a reasonable basis to believe has been lawfully made available to the general public by—government records; widely distributed media; a website or online service; or a disclosure that is required to be made by law.	15

can be helpful for policy analysts because, if a term has a high number of dependencies, this suggests the term is structurally integral for a bill and may be an important area to focus resources. For instance, in the APRA, the term "sensitive covered data" has the highest out-degree, serving as an umbrella term that ties together many other terms defined in the bill and critically outlining the bill's scope. Policy staff analyzing the bill might want to focus on this definition to ensure it is robust and has appropriate coverage, especially given that the definition may be a source of debate amongst those weighing in on the bill.

In contrast, a low out-degree or an out-degree of zero for a given term means its definition relies on few other terms in the bill. The interpretation of a low out-degree depends on the context: In some cases, a low out-degree can indicate that a term is unclear or underspecified. For example, in the APRA, the term "clear and conspicuous" has an out-degree of zero. As scholars have highlighted,

TABLE 5

Possible interpretations of a term's relationships to other definitions, using examples from the American Privacy Rights Act of 2024 (APRA).

	Low	High
Out-degree: Number of times a term refers to other terms	Terms that may be underspecified or may be clear without further context	Terms that rely on many other terms and may be structurally important in bringing together other terms
	Example: "clear and conspicuous" in APRA	Example: "sensitive covered data" in APRA
In-degree: Number of times a term is referred to by other terms	Terms that are less relied on by others in the bill and may be clear without further context	Terms that are relied on by many other terms and may be structurally important
	Example: "parent" in APRA	Example: "covered data" in APRA

sometimes these definitions may be vague by design and intended to be clarified through other policy vehicles as the law is implemented. But ironically, a recent study found that in the context of laws related to AI, definitions of "clear and conspicuous" notices or disclosures are sometimes not so clear. By identifying terms with a low out-degree, policy staff can evaluate whether those definitions should be clarified, including potentially with references to other terms in the bill.

In other instances, a low out-degree may correspond to a term that is generally clear without further context or reliance on other definitions. For example, in the APRA, the terms "commission" and "state" both have an out-degree of zero. The definition of "commission" makes it clear that the use of the word within the bill refers to the Federal Trade Commission, and the definition of "state" indicates that it refers to "each of the 50 States, the District of Columbia, the Commonwealth of Puerto Rico, the Virgin Islands of the United States, Guam, American Samoa, and the Commonwealth of the Northern Mariana Islands."

As with analyzing cycles, ascribing a qualitative meaning to a term's importance based on its degree depends on the context, including the structure and nature of the bill in question, as well as examining a term's in-degree and out-degree together. But by identifying how terms rely on each other and are concentrated together within a bill, we may ultimately be able to operationalize network analysis metrics — such as a node's in-degree and out-degree — as indicators of a term's complexity and importance in a bill. For example, **Table 5** provides one possible interpretation of the frequency with which terms rely on each other in a bill. Policy staff could reference and apply this kind of heuristic to the metrics produced from a bill's definition to help focus their time and attention when analyzing AI policy.

3.4. Further potential for applying graph theory to legislative data

There are many further possibilities for applying graph theoretical methods to legislative data, building on the methods described in this report. For instance, expanding the analysis of cycles beyond binary detection, we could propose and test computational methods for qualitatively classifying different types of cycles in definitions — such as a notion of definitional dependency to identify when an edge between term x and term y indicates that x depends on y, as opposed to merely referencing it — and applying inter-bill analysis tools to connect definitional graphs of multiple bills. We could also consider and potentially quantify notions of exclusions or missingness in definitional dependencies using these graphs. For example, are there terms that should reference each other but don't?

More broadly, future work could apply more sophisticated graph-theoretic analysis that ascertains which definitions are most "central" or "load-bearing" within a text. For example, these could be definitions that both rely on earlier definitions and act as key elements in more "downstream" definitions, in such a way that eliminating these definitions would destroy the structure of the graph. This approach relates to ideas of concept "centrality" within the network analysis community.

4. A Call for Standardized Formats and Structures for Bill Texts

We encourage researchers to work with policy staff to create standardized formats and structures for bill texts across jurisdictions. We identify key challenges when computationally parsing legislation that motivate this need for standardization and outline potential paths forward.

4.1. Challenges of Parsing Legislative Documents

First, legislative documents often come in **inconsistent file formats**, ranging from structured XML or HTML to semi-structured DOCX and unstructured text in PDF. This inconsistency often poses challenges to automatic text extraction and complicates further analysis of the text.

Legislative documents also often have **varied structures across jurisdictions**, even when a structured file format like XML is used. For example, there is often variability in the formatting of headers and footers, section names, definitions, line numbers, and spacing. Using computational techniques such as machine learning can help parse, clean, and infer structures, but due to the high variability in document formatting, these techniques can still produce errors that may propagate to downstream computational analysis.

Third, legislative documents **annotate changes to texts inconsistently** across jurisdictions. Legislators use two key types of annotations: (1) *amendments* to indicate proposed changes to an existing law, and (2) *version tracking* to indicate changes to a previous version of the same legislative proposal. While amendments to existing laws can represent a significant policy shift, version tracking is important to understand the legislative process. However, amendments and versioning are sometimes annotated in the same manner, removing the nuances between them. There is also variability in whether and how these annotations are shown across and sometimes even within jurisdictions.[13] Inconsistency in the file formats can further complicate the ability to parse these annotations. For example, annotations in the form of colored text and strikethroughs are typically visible as tags in an HTML file, but for a PDF file, automated parsing tools may not capture these annotations.

Finally, legislative documents typically have **insufficient references** that link either internally to defined terms and sections or externally to existing laws, statutes, and agencies. These references can facilitate faster and more reliable construction of graphs that map the relationships across definitions as well as legislation. Inconsistency in the file formats and structures complicates the ability to parse references. For example, if a bill is only available in PDF format, it is unlikely to have these references. Even if a bill is available in HTML or XML format, jurisdictions encode these references in different ways, if at all.

4.2. Proposed Solution for Standardizing Legislative Documents

To address these three challenges, we recommend researchers work with policy staff to establish a consistent file format and a standardized structure for the file content. We propose establishing the following standards for legislative documents across jurisdictions:

File format: Standardize the usage of user-friendly DOCX files for legislative documents. Using a consistent structure for the content in the DOCX, as outlined next, could facilitate a direct backend conversion of a bill to a more machine-readable file format like XML, as well as a more human-readable format such as HTML.

Document structure: Establish a pre-defined set of section categories (e.g., "Definitions," "Requirements," and "Enforcement") and a format for how terms are defined in legislative documents. Standardizing the structure of legislative documents would minimize text parsing errors and allow for more appropriate document segmentation, facilitating downstream computational analysis.[14]

Annotations: Establish standards for the format and usage of annotations, for both amendments and version tracking. A version control system (VCS) like Git could facilitate annotating version changes to legislative texts, as the VCS would render them automatically. This would also allow policy staff to inspect differences between any pair of bill versions, not just consecutive ones.

References: Standardize creating tags or hyperlinks for references in a given bill to other sections and terms in the bill as well as other bills. Consistent internal references, such as between terms, would facilitate *intra-bill* analysis, including more reliable and automatically generated graphs of definitions. Consistent external references, such as amendments to existing state statutes, would enable *inter-bill* analysis, including generating reference graphs across proposed legislation and existing law. Consistent external references to existing entities, such as federal agencies, would also aid in analyzing the relevant regulatory or oversight bodies for a given bill and across bills.

We identify resources and templates that could serve as good starting points for standardization efforts across jurisdictions

Researchers and policy staff can iteratively update these standards to accommodate new bill structures and jurisdiction- and chamber-specific needs.

4.3. Building on Existing Resources for Standardizing Legislative Documents

We next discuss how existing resources can serve as a starting point for making this proposed shared standard a reality across different jurisdictions. First, researchers and policy staff can collaborate with established groups, such as the Congressional Data Coalition and the international LegalXML community, to guide these efforts. They can also build upon existing tools for standardization. For example, Congress and several states — including California, Washington, and Texas — already provide bill texts in structured file formats like HTML and XML through their websites, such as Congress.gov. Researchers and policy staff can adapt and extend the United States Legislative Markup XML schema, which standardizes file formats, document structures, and annotations for congressional legislative documents. They can also leverage open-source tools such as Congress.dev and GovTrack, which display differences between bill versions and across bills, to develop annotation standards. Finally, applying NLP methods can help standardize the structure and format of legislative texts, although these tools should be supplemented with human review to ensure accuracy.

FIGURE 8

Proposed changes to the AI Civil Rights Act of 2024 to facilitate intra-bill definition analysis.

Text: The terms “collect” and “collection”, with respect to personal data, mean […]

Current XML

```xml
<text>
  The terms <term>collect</term>
  and <term>collection</term>,
  with respect to personal data,
  mean […]
</text>
```

Proposed XML

```xml
<definitions>
  <terms>
    <term>collect</term>
    <term>collection</term>
  </terms>
  <text>
    The terms <term>collect</term>
    and <term>collection</term>,
    with respect to <term>personal data<term>,
    mean […]
  </text>
</definitions>
```

The definition of the term “collect/collection” is shown at the top and rendered in “Text”. The current XML, sourced from Congress.gov, is shown on the left and the proposed XML on the right.

Benefits of Proposed XML

- Encodes **definition-specific section**
- Encodes **defined terms** including synonyms, aliases, and acronyms
- Encodes **internal references** to other terms like *personal data* within the bill

⟶ Facilitate **graph construction** of terms *within* the bill

FIGURE 9

Proposed changes to the AI Civil Rights Act of 2024 to facilitate inter-bill reference analysis.

Text: The term “consequential action” means […] as determined by the Federal Trade Commission through rules promulgated pursuant to section 553 of title 5, United States Code.

Current XML

```xml
<text>
  The term <term>consequential action</term>
  means […] as determined by
  the Federal Trade Commission
  through rules promulgated pursuant to
  section 553 of title 5, United States Code.
</text>
```

Proposed XML

```xml
<definitions>
  <terms>
    <term>consequential action</term>
  </terms>
  <text>
    The term <term>consequential action</term>
    means […] as determined by
      <ref type="existing-entity" link=...>
        the Federal Trade Commission
      </ref>
    through rules promulgated pursuant to
      <ref type="existing-code" link=...>
        section 553 of title 5, United States Code.
      </ref>
  </text>
</definitions>
```

Part of the definition of “consequential action” is shown at the top. The current XML, sourced from Congress.gov, is shown on the left and the proposed XML on the right.

Benefits of Proposed XML

- Encodes **definition-specific section**
- Encodes **defined terms**
- Encodes external references to existing **entities** like the *Federal Trade Commission*
- Encodes external references to existing **codes and statues** like the *United States Code*

⟶ Facilitate **graph construction** *across* different documents, as well as graph *between* documents and entities

4.4. Case Study: Standardizing the AI Civil Rights Act of 2024

To illustrate how these existing standards can be strengthened, we provide examples of proposed modifications to the AI Civil Rights Act of 2024, using the bill's original XML file obtained from Congress.gov.

We first focus on the definition of the terms "collect" and "collection" in the AI Civil Rights Act of 2024, as shown in **Figure 8**. We propose modifying the original XML of this definition in three key ways: (1) clearly indicate that this definition is part of a "Definitions" section to help locate it in the AI Civil Rights Act of 2024 (highlighted in blue), (2) explicitly encode the terms "collect" and "collection" that are defined together (highlighted in red), and (3) tag internal references to other terms in this definition (highlighted in purple). These changes can facilitate intra-bill analysis, such as generating a graph of the terms defined in the AI Civil Rights Act of 2024.

We next examine the definition for the term "consequential action" in the AI Civil Rights Act of 2024, as illustrated in **Figure 9**. In addition to some of the previously mentioned modifications, we show how the original XML of this definition could be improved with external references. Specifically, we propose encoding the references to the Federal Trade Commission (FTC) (highlighted in green) and the United States Code (highlighted in yellow) in the definition. These changes can enable inter-bill analysis across different documents and entities, such as identifying the FTC as a relevant regulatory agency for the AI Civil Rights Act of 2024 and constructing a graph of the references between the AI Civil Rights Act of 2024 and existing codes.

5. A Call for Multilingual AI Policy Analysis

In addition to our recommendation for standardized bill texts, we encourage researchers and advocates to incorporate a multilingual perspective when analyzing AI legislation introduced in regions that, due to the history and ongoing reality of U.S. imperialism, are under U.S. jurisdiction.

Due to capacity and language skills, we were limited to designing a tool that only handles AI policy written in English. As a result, we excluded legislative texts written in other languages, which future work might analyze through a multilingual approach. For example, Puerto Rico's legislature has introduced over 20 bills related to AI, all written in Spanish. These bills cover a range of topics, including creating a government AI officer role, prohibiting AI-generated explicit content, and regulating the use of AI in the education system. Hawai'i's legislation is sometimes written partially or fully in Hawaiian, although Hawai'i's AI-related bills that we are aware of have been published in English.

Using computational methods to analyze such non-English AI policy should leverage **language technology tailored to specific languages** and involve native speakers throughout the process. Mainstream language models built by prominent tech companies often perform poorly in non-English languages, rely on machine-translated instead of human-translated text, and are not developed with the communities they aim to serve. In contrast, language-specific AI research groups, such as AmericasNLP for Indigenous American languages, focus on building AI tools in their own languages that incorporate social and cultural context. Some computational methods, such as converting legislative texts into machine-readable formats and extracting effective dates, may be more transferable across languages and may not require deep language expertise. However, other methods, like topic modeling and comparing sentences across bills, depend heavily on social context and should use tools created by language-specific researchers.

Computational AI policy analysis should also leverage **language technology tailored to legal contexts** and involve regional AI policy experts. Mainstream language models are typically not tailored to tasks important in legal contexts, such as following instructions, especially in non-English languages. However, language-specific researchers and practitioners have developed models tailored to legal texts in Spanish, such as Legal-ES, Narralegal, and Modelo de Español Legal.

Translation and interpretation by people are context-dependent, and each context requires specialized skills. For instance, U.S. federal courts outline certification requirements for interpreters working in legal settings, recognizing that there are *languages within languages* and that communication in legal contexts differs significantly from communication in other contexts like medicine. Similarly, any potential use of language technologies in specialized contexts must recognize and incorporate this nuance. Integrating such language- and legal-specific knowledge in computational approaches is essential to accurately understand, uplift, and learn from the diverse approaches to AI policy in regions under U.S. jurisdiction.

6. Conclusion

Throughout this report, we have demonstrated the potential utility of applying data-driven approaches to analyze AI and ADS legislation, specifically using graph theory and natural language processing methods. We applied these different techniques in isolation but future research could combine them to answer more nuanced questions such as: Do *Synthetic Content & Generative AI* bills tend to model each other more than bills on other topics do? Do *Consumer Protection & System Regulation* bills tend to have more well-connected and internally reliant definition sections than *AI & Law Enforcement* bills do, as the latter may rely on more established terms? Future work might also involve developing standardized formats and structures for bill texts, and analyzing AI and ADS policy with a multilingual perspective, as our calls to action highlight.

Computational methods can be useful to better understand the AI and ADS legislative landscape, yet we also emphasize that these tools should not replace review from legal experts, civil rights groups, legislators, and researchers. Rather, computational analysis can offer such experts a toolkit for inspecting the current legislative landscape at different scales (e.g., inter-bill versus intra-bill). We hope this work helps policy staff better understand and strengthen the growing landscape of legislation on AI technologies.

Acknowledgements

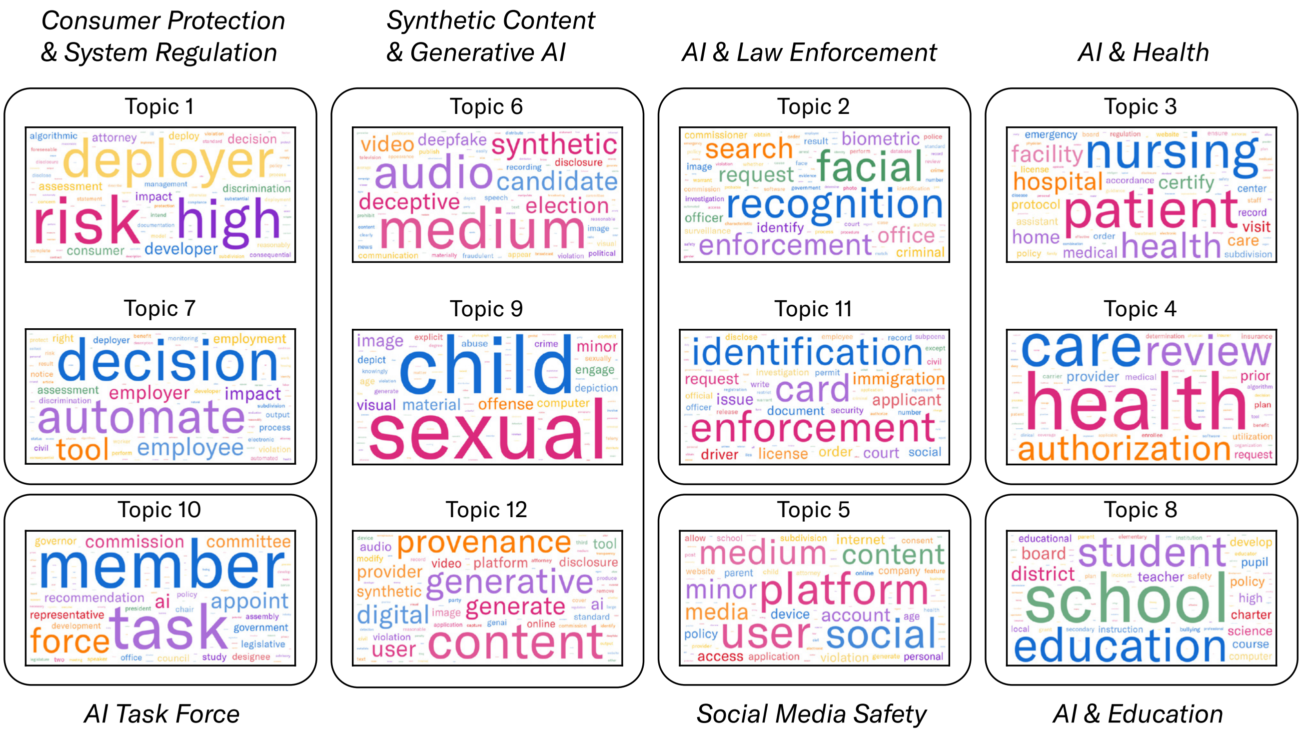
CNTR would like to acknowledge: funding source from Media Democracy Project, bill data source and API from the Open States project operated by Plural Policy, data storage from DagsHub, and computing resources from Brown CCV. Other people involved in different parts of the CNTR AISLE AI Legislative mapping project at CNTR: Sasa Jovanovic, Mahir Arora (involved in sectioning), Nora Cai (involved in data collection), Timothy Fong (involved in data collection), Rachel Kim, Fern Tantivess.

The ACLU Tech Team would also like to acknowledge its partnership with the Patrick J. McGovern Foundation.

Appendix

FIGURE A1

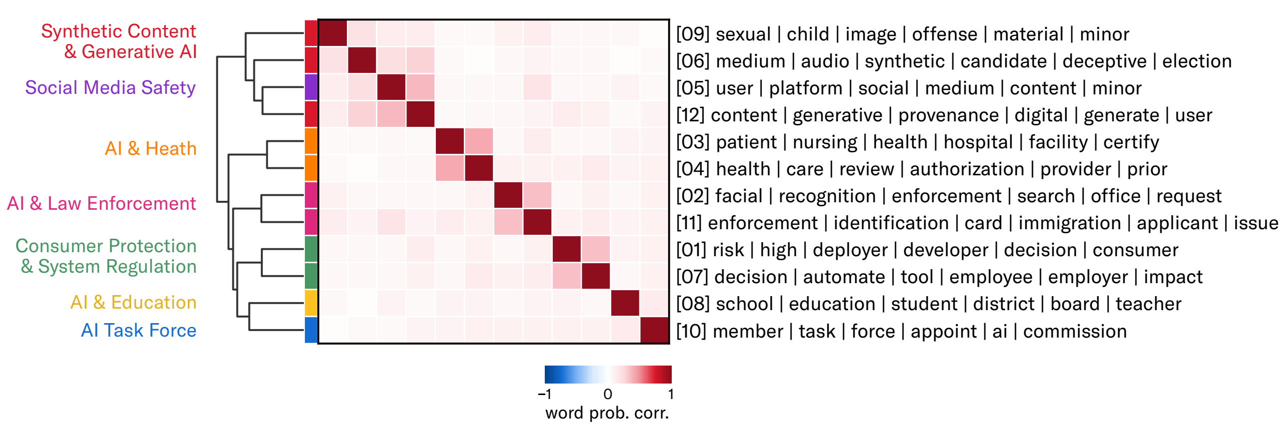
Legislative clusters from topic modelling



Manual post-hoc groupings of topics after topic modelling. Each word represents a topic - the size of each word corresponds to their importance and representation in the corresponding topic. Words are colored randomly for visibility. Manual groupings were done post-hoc to reduce 12 topics to 7 clusters for ease of analysis. See also **Figure A2** for clustering analysis with topic probability representation that eases the manual groupings.

FIGURE A2

Topic hierarchical clustering via word probability representation



Each topic has a word representation vector. Their pairwise correlation matrix is shown here with hierarchical clustering of these topics. This analysis is used to inform the final manual groupings in the main text. The colors under the hierarchical clustering are the same as the finalized thematic legislative clusters shown in **Figure 3**.

TABLE A1

Manual post-hoc groupings of topics to legislative clusters in topic modelling analysis

Cluster	Topic	Example bills per topic with probability >= 70%
Consumer Protection & System Regulation	[01] risk | high | deployer | developer | decision | consumer	• GA SB 167 (2025_26): Commerce and Trade; private entities that employ certain AI systems to guard against discrimination caused by such systems; provide • MD HB 1331 (2025): Consumer Protection - Artificial Intelligence • IA HSB 294 (2025-2026): A bill for an act relating to artificial intelligence, including the use of artificial intelligence to create materials related to elections and protections in interactions with artificial intelligence systems, and making penalties applicable. • VT H 341 (2025-2026): An act relating to creating oversight and safety standards for developers and deployers of inherently dangerous artificial intelligence systems • RI HB 7786 (2024): AN ACT RELATING TO COMMERCIAL LAW – GENERAL REGULATORY PROVISIONS – AUTOMATED DECISION TOOLS
	[07] decision | automate | tool | employee | employer | impact	• NY S 2277 (2023-2024): Enacts the “digital fairness act” • US S 2892 (118): Algorithmic Accountability Act of 2023 • IL HB 5116 (103rd): AUTOMATED DECISION TOOLS • OK HB 3835 (2024): Technology; title; Ethical Artificial Intelligence Act; deployers; developers; algorithmic discrimination; attorney general; effective date. • US S 1865 (118): TAG Act
AI Task Force	[10] member | task | force | appoint | ai | commission	• OR HB 4153 (2024R1): Relating to artificial intelligence; declaring an emergency. • US S 1356 (118): ASSESS AI Act • NV SB 165 (82): Revises provisions relating to businesses engaged in the development of emerging technologies. (BDR 18-878) • US S 2293 (118): AI LEAD Act • CT HB 5047 (2025): AN ACT CREATING A TASK FORCE TO STUDY ARTIFICIAL INTELLIGENCE AND THE STATE WORKFORCE.
Synthetic Content & Generative AI	[06] medium | audio | synthetic | candidate | deceptive | election	• NV AB 73 (83): Establishes requirements for certain communications relating to an election. (BDR 24-487) • OH HB 410 (135): Regulate dissemination of deepfake media to influence an election • LA SB 97 (2024): POLITICAL CAMPAIGNS: Regulates the use of deep fakes and artificial intelligence technology in political advertising. (gov sig) • KS SB 375 (2023-2024): Prohibiting the use of generative artificial intelligence to create false representations of candidates in election campaign media or of state officials. • NH HB 1596 (2024): requiring a disclosure of deceptive artificial intelligence usage in political advertising.
	[09] sexual | child | image | offense | material | minor	• CA AB 1856 (20232024): Disorderly conduct: distribution of intimate images. • CT SB 1440 (2025): AN ACT CONCERNING UNAUTHORIZED DISSEMINATION OF INTIMATE IMAGES THAT ARE DIGITALLY ALTERED OR CREATED THROUGH THE USE OF ARTIFICIAL INTELLIGENCE. • GA HB 1361 (2023_24): Crimes and offenses; distribution of computer generated obscene material depicting a child; prohibit • WA SB 5094 (2025-2026): Concerning sexually explicit depictions of minors. • OH SB 163 (136): Regards AI images, simulated child porn, replica identity fraud
	[12] content | generative | provenance | digital | generate | user	• US S 4674 (118): Content Origin Protection and Integrity from Edited and Deepfaked Media Act of 2024 • MA H 81 (194th): An Act relative to artificial intelligence disclosure • NY A 6540 (2025-2026): Requires generative artificial intelligence providers to include provenance data on certain content made available by the provider • FL SB 702 (2025): Provenance of Digital Content • MA HD 1861 (194th): An Act regulating provenance regarding artificial intelligence

AI & Law Enforcement	[02] facial | recognition | enforcement | search | office | request	• MT HB 532 (2025): Generally revise laws related to abuse, neglect, and exploitation of incapacitated persons and vulnerable adults • US HR 6092 (118): Facial Recognition Act of 2023 • NY A 10625 (2023-2024): Relation to the regulation of the use of artificial intelligence and facial recognition technology in criminal investigations • MT HB 267 (2025): Revise DUI laws related to enacting Bobby's law • MT LC 1312 (2025): Prohibit facial recognition technology at traffic lights
	[11] enforcement | identification | card | immigration | applicant | issue	• IL SB 2649 (103rd): REPEAL ILLINOIS TRUST ACT • IL HB 3882 (103rd): STANDARD ID-DOCUMENTATION • IL SB 3596 (103rd): IMMIGRATION ENFORCEMENT ACT • NY S 8390 (2023-2024): Relates to the admissibility of evidence created or processed by artificial intelligence • IL HB 4495 (103rd): COUNTY RESIDENCE ON ID/LICENSE
AI & Health	[03] patient | nursing | health | hospital | facility | certify	• VA HB 664 (2024): Abortion; born alive infant, treatment and care, penalty. • VA SB 925 (2023): Patient visitation; visitation from clergy members during declared public health emergency. • VA SB 392 (2024): Hospitals; emergency departments to have at least one licensed physician on duty at all times. • VA HB 87 (2024): Hospital regulations; patient drug testing. • VA HB 886 (2024): Certified nursing facilities; administrative sanctions, facilities subject to minimum standards.
	[04] health | care | review | authorization | provider | prior	• US HR 206 (118): Healthy Technology Act of 2023 • US S 923 (118): Better Mental Health Care for Americans Act • TX SB 1411 (89R): Relating to the use of artificial intelligence-based algorithms by health benefit plan issuers, utilization review agents, health care providers, and physicians. • RI HB 5172 (2025): AN ACT RELATING TO INSURANCE – THE TRANSPARENCY AND ACCOUNTABILITY IN ARTIFICIAL INTELLIGENCE USE BY HEALTH INSURERS TO MANAGE COVERAGE AND CLAIMS ACT • TX HB 4455 (89R): Relating to the use of artificial intelligence by health care providers.
AI & Education	[08] school | education | student | district | board | teacher	• TN HB 933 (114): Education, Dept. of - As introduced, requires the department to establish and administer a three-year artificial intelligence weapons detection system grant pilot program to award grants to eligible LEAs for the purchase of artificial intelligence weapons detection systems for schools without an artificial intelligence weapons detection system. - Amends TCA Title 49. • IL SB 1677 (104th): SCH CD-TEACHER EVALUATION PLAN • MO HB 2612 (2024): Establishes an educational technology impact advisory council to review the use of technology in schools • CA AB 2652 (20232024): State Department of Education: artificial intelligence working group. • TX HCR 53 (89R): Congratulating the Canyon High School FFA Ag Issues Team on winning first place in the 2024 Texas FFA State Agricultural Issues Forum.
Social Media Safety	[05] user | platform | social | medium | content | minor	• MA HD 3070 (194th): An Act promoting safe technology use and distraction-free education for youth • NC SB 514 (2025): Social Media Control in IT Act. • MN SF 1857 (2025-2026): Minor access to chatbots for recreational purposes by persons prohibition provision • IL HB 3943 (103rd): SOCIAL MEDIA MODERATION • OK SB 885 (2025): Social media; creating the Safe Screens for Kids Act. Effective date.

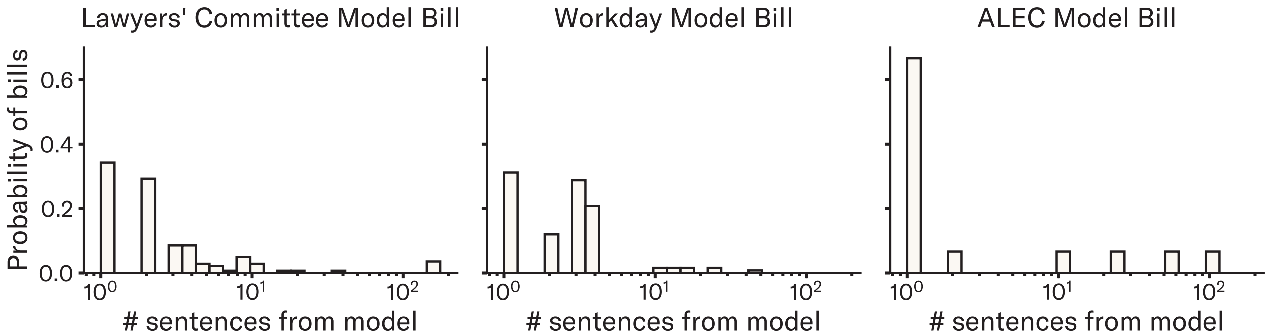
FIGURE A3

Maps of topics

Each panel shows the bill count for each legislative cluster across the United States, with congressional bills indicated in the lower right corner (*US*). Note color bars have different maximums for different clusters to highlight state differences per cluster.

FIGURE A4

Distribution of number of model bills' sentences that share high similarity with legislative bills.



For a given model bill, its text (panels) is segmented into sentences, and only sentences with at least 10 words are then compared to legislative texts using partial fuzzy ratios (see main text for methodology). High-similarity of a sentence from a model bill means such a sentence reaches the 80% partial ratio threshold when compared to a given legislative bill. The distribution is over the 1804 bills analyzed in this report, in semilog scale. The threshold of minimum 10 sentences (in **Figure 5**) is used to determine whether a legislative text shares high similarity with a model bill. We choose this threshold to be (a) consistent across all three templates, and (b) stricter for the longer Lawyers' Committee Model Bill.

Endnotes

1 We selected the start of 2023 for two key reasons: (1) ChatGPT was released at the end of 2022 and (2) many states did not hold regular sessions the year before. We stopped collecting data in mid-April 2025 to focus on finalizing our analyses.

2 The full list of keywords was: "artificial intelligence", "automated decision making", "automatic decision making", "decision making tool", "automated decision tool", "automatic decision tool", "automated decision system", "automatic decision system", "automated final decision", "automatic final decision", "face recog", "facial recog", "voice recog", "iris recog", "gait recog", "genAI", "gen-AI", "generative AI", "generative tech", "generative model", "generative artificial", "machine learning", "deep learning", "computer vision", "natural language process", "language model", "ChatGPT", "Chat-GPT", "pre-trained transformer", "stable diffusion", "AI task force", "AI advis", "AI audit", "AI generate", "AI snoop", "deep fake", "synthetic media", "frontier model", "digital assistant", "chat bot", "virtual assistant", "software agent", "virtual agent", "embodied robot", "foundation model", "foundational model", "open source AI", "agentic AI", "LLM", "LLMs", "Information Technology Act"

Note 1: We replace spaces, if present, in these key terms with the regular expression "*\s*[^\w]?\s*" to accommodate small variations, e.g. instead of only detecting "chat bot", such change would allow detecting "chatbot" and "chat-bot".

Note 2: We do not use the exact acronym "AI" for multiple reasons, including that (a), "ai" is a word or word part in the Hawaiian language that sometimes appears in bills introduced in the Hawai'i state legislature, (b) "AI/AN" is often used as an acronym for "American Indian and Alaska Native", and (c) "AI" may appear as an index (e.g. "AA", "AB")

Note 3: We included the „Information Technology Act" to detect certain bills (e.g. Utah SB 131) that did not end up in the bill search initially, but that we knew were AI or ADS bills from our own research. We then removed this term during preprocessing with anchor keywords.

Note 4: After collection, we added more keywords to be used as anchors: "AI assistant" and "digital depiction". We also used a regular expression to detect and count special cases where occurrences of recognition technologies were concatenated. For instance, "*facial, voice, iris, and gait-recognition software*" in Virginia HB 1496 should be counted as 4 occurrences instead of 1.

3 The actual regular expression is "*automat\w+ decision*" to capture variations such as "automated", "automatic"

4 Specifically, these are bills with either (a) at least 4 matched keywords (counting duplicating keywords) *or* (b) where the logarithmic ratio between the number of matched keywords and the word count of the text (after trimming based on anchor words) is at least -2. By default, bills with "artificial intelligence" or "automat(ed|ic) decision" would be included, even if these terms only occur once and do *not* appear together.

5 We define these as any bills whose title includes any of "fiscal", "budget", or "appropriation."

6 Though not included in this analysis, Puerto Rico's (PR) legislatures have introduced and are advancing numerous pieces of legislation related to AI and automated decision-making systems (see, e.g., PR HJR 68 (2025), PR S 68 (2025), PR S 622 (2025)).

7 See, for instance, Artificial Intelligence Index Report 2025, published by the Stanford University Center for Human-Centered Artificial Intelligence, pg. 337.

8 For example, in New York A-3593, a bill largely focused on privacy and data-protection titled "*Relates to enacting the NY privacy act*", only the initial definition section and one other section mention automated decision-making.

9 We trimmed bills with more than 5000 words by centering on anchor keywords. For these bills, we keep the initial 1000 characters of the bill, and for every anchor keyword, we also keep the 1000 characters before and 2000 characters after the anchor keyword. Future work could explore more robust approaches, such as using sentences or paragraphs, instead of characters, surrounding anchor words to trim.

10 For each bill, this method produces a set of probabilities that the given bill belongs to each of the 12 topics, and we set a threshold for topic probability — classifying a bill as belonging to a topic if it has a probability of at least 25% for the topic – to strike a balance between bills with multiple topics and confidence of topic modelling results. In the future, we could consider splitting a bill into smaller chunks, then run topic modelling results on all chunks instead.

11 If two legislative texts, Bill A and Bill B, were exactly identical, the corresponding heatmap would have a dark blue diagonal line running from the top left to the bottom right of an otherwise white graph, as the first sentence of Bill A would match the first sentence of Bill B (and does not match any other sentences), the second sentence of Bill A would match the second sentence of Bill B, and so on.

12 Note that we do not know whether or when the text of the Workday Model Bill was shared with legislators in various states (as *The Record* article reports), only that the The Record article in March 2024 made the text of the template public and revealed that many state bills shared high similarity with the template. As a result, we cannot explain why there were bills in 2023 sharing similarity with this template. It might be the case that all or a subset of these bills influenced the drafting of the template in the first place, or that internal sharing of the template with legislators influenced these state bills. This uncertainty and lacking transparency is also a demonstration of the difficulty in determining causality with regard to policy diffusion.

13 For example, California SB-1047 proposes but does not annotate amendments to existing law, yet annotates changes across versions of the bill itself. In contrast, California SB-354 annotates amendments to existing law in the bill's first version as well as changes across versions of the bill.

14 For example, thematic analysis would benefit greatly from input legislative data that is free of the potential errors and variability that we usually encounter during text extraction and preprocessing from PDF files. Furthermore, with standardized structures, robust document segmentation would allow more robust multi-topic assignment for a given piece of legislation.