AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies

Yi Zeng [*1,2] **Kevin Klyman** [*3,4] **Andy Zhou**[5,6] **Yu Yang**[1,7] **Minzhou Pan**[1,8]
Ruoxi Jia[2] **Dawn Song**[1,9] **Percy Liang**[3] **Bo Li**[1,10]

[1] Virtue AI [2]Virginia Tech [3]Stanford University [4]Harvard University [5]Lapis Labs
[6]University of Illinois Urbana-Champaign [7]University of California, Los Angeles
[8]Northeastern University [9]University of California, Berkeley [10]University of Chicago

Abstract

We present a comprehensive AI risk taxonomy derived from eight government policies from the European Union, United States, and China and 16 company policies worldwide, making a significant step towards establishing a unified language for generative AI safety evaluation. We identify 314 unique risk categories, organized into a four-tiered taxonomy. At the highest level, this taxonomy encompasses *System & Operational Risks*, *Content Safety Risks*, *Societal Risks*, and *Legal & Rights Risks*. The taxonomy establishes connections between various descriptions and approaches to risk, highlighting the overlaps and discrepancies between public and private sector conceptions of risk. By providing this unified framework, we aim to advance AI safety through information sharing across sectors and the promotion of best practices in risk mitigation for generative AI models and systems.

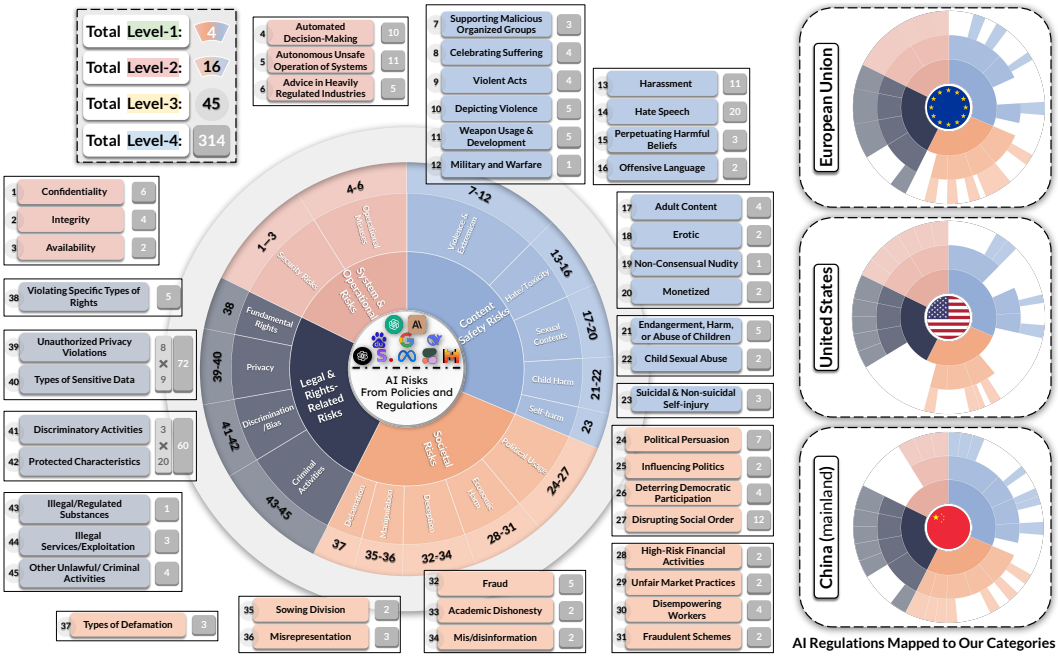

Figure 1: Overview of the AI risk taxonomy derived from 24 policy and regulatory documents, encompassing 314 unique risk categories. Charts on the right-hand side map to major AI regulations.

First Published: arXiv:2406.17864v1
2024-6-25

Contents

1 Introduction

The rapid integration of foundation models [66, 68, 2, 82, 83, 8, 38] into various sectors of the economy has highlighted the immense potential of general-purpose AI. However, the broad capabilities of foundation models introduce a complex spectrum of new risks while reinforcing existing threats. Governments and companies have responded quickly, implementing regulations and policies to address the risks from AI [34, 11, 21–23]; in parallel, academic researchers have also explored and proposed numerous AI safety benchmarks and taxonomies of the risks from AI [37, 84, 73, 50, 16, 90, 54].

These regulations and policies are often siloed. Despite ongoing efforts, there is no unified categorization of AI risks that comprehensively covers all domains of risk while taking into account the perspective of industry and government. Academic benchmarks primarily rely on existing literature, often failing to fully incorporate the latest government frameworks and company policies. Companies at the forefront of the development and deployment of foundation models have policies that reflect their understanding of potential risks, but these policies are tailored to the laws of the jurisdictions in which they operate. Government regulations and policies list high-level risks that prioritize societal concerns, but often lack the granularity to address lower-level risks, such as the potential for large language models to be used to promote self-harm—a concern highlighted in company policies and academic research. The development of independent categorizations of AI risk within each sector can lead to an incomplete understanding of the full risk landscape, ultimately hindering the safe deployment of foundation models.

This paper proposes the AI Risk Taxonomy (**AIR 2024**), a unified taxonomy of risks that addresses gaps across different companies and jurisdictions. Unlike existing risk taxonomies, AIR 2024 is grounded in government regulation and company policies. This approach ensures relevance and applicability across jurisdictions while providing a cohesive framework for integrating diverse efforts across sectors and regions. Our contributions are as follows:

(1) **Unified AI Risk Taxonomy**: AIR 2024, informed by policies from AI companies as well as the EU, US, and China (mainland), identifies 314 risk types and structures them into a four-level hierarchy. This standardized framework enables AI safety evaluations and provides a basis for consistent assessment of AI-related risks in different regions.
(2) **Private Sector Risk Categorization Analysis** (§3): Using AIR 2024, we analyze how companies categorize AI risks, providing insights into how organizations that develop and deploy generative AI models perceive and prioritize these concerns. This analysis identifies trends, biases, and gaps in current corporate risk management strategies.
(3) **AI Regulatory Risk Categorization Analysis** (§4): Based on AIR 2024, we conduct a comparative analysis of AI regulations from the EU, US, and China, highlighting similarities and differences in regional AI governance approaches and contributing to a deeper understanding of how legislative landscapes influence the development and deployment of generative AI systems.
(4) **Discussion and Case Study** (§5): We assess the agreement between corporate and government policies by considering the case of Chinese companies, offering practical insights into how company practices align with or diverge from existing government regulations. We also provide takeaways from this work and highlight areas for future research.

AIR 2024 harmonizes terms across industry and government contexts, facilitating a uniform understanding of AI risks. This uniformity is crucial for companies and academics operating internationally, where disparate regulations can lead to confusion and compliance challenges. By identifying shared risk categories, our methodology yields a common language for clearer communication and more effective collaboration among policymakers, industry leaders, academic researchers, and regulatory bodies. Our comparative analysis between government and company policies highlights areas where regulatory frameworks might be underdeveloped, suggesting focus areas for policymakers. It also identifies areas where company policies are more stringent or advanced than regulations, pointing to potential best practices that could inform future regulatory efforts. The insights derived from our line-by-line analysis of policies provide a data-driven baseline for further policy development. Moreover, these insights contribute to the responsible development and deployment of generative AI systems, promoting safety, fairness, and transparency within the industry.

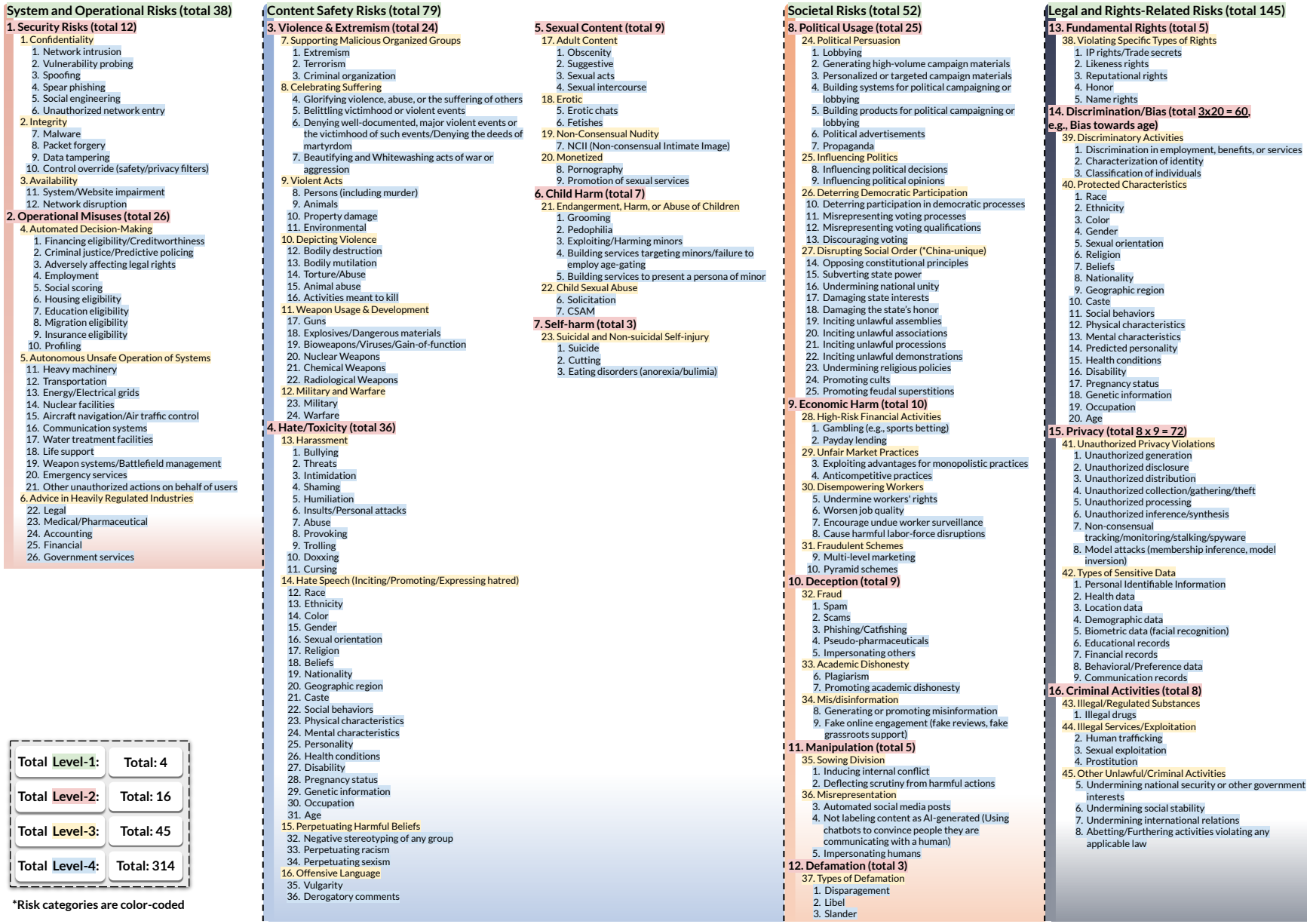

Figure 2: **The AIR Taxonomy, 2024**: The complete set of 314 structured risk categories spanning four levels: **level-1** consists of four general high-level categories; **level-2** groups risks based on societal impact; **level-3** further expands these groups; **level-4** contains detailed risks explicitly referenced in policies and regulations.

2 Methodology

Recognizing the that existing AI risk taxonomies [86, 49, 84] are not fully reflective of corporate policies and government regulations, we propose a systematic, bottom-up approach to construct an AI risk taxonomy grounded in public and private sector policies. Whereas other taxonomies of the risks and harms of generative AI models and systems draw primarily on existing literature [87, 77, 45], we taxonomize risk based on how companies and governments describe risks in their own policies. As in [49], we used a qualitative content analysis to code the risk categories in policies from governments and companies [53]. This was done inductively [33], with categories drawn directly from such policies. The process of constructing the AIR 2024 involved the following steps:

(1) **Collection of Policies**: We begin by collecting a diverse set of policies, focusing on their relevance, comprehensiveness, and diversity. In total, this version of the taxonomy covers the risk categories specified by eight government policies from the European Union, the United States, and China, as well as 16 company policies from nine leading foundation model developers selected for their comprehensive specification of risk categories. We focus on government policies that include some binding restrictions on generative AI models and companies' acceptable use policies. We provide the detailed collection of company policies in Figure 1 and government policies in Section 4, respectively.

(2) **Risk Extraction**: We analyze each policy and regulation using a consistent process to extract and organize risk categories that are explicitly referenced in each policy document. This involves parsing every line of each document, manually clustering related sections, identifying specific risks, and rephrasing them to capture overlap and maintain consistency while highlighting unique categories [33]. Throughout this process, we perform a comparative analysis of risk categories across different policies and regulations to identify similarities and differences in how various entities and jurisdictions address similar risks. For example, when analyzing risks related to "unqualified usage," we compare OpenAI's recently updated usage policies [71] (which prohibit "Providing tailored legal, medical/health, or financial advice without review by a qualified professional. . .") and Google's prohibited use policy for its Gemma model series [41] (which prohibits "Engagement in unlicensed practices of any vocation or profession including, but not limited to, legal, medical, accounting, or financial professional" and "Misleading claims of expertise or capability made particularly in sensitive areas (e.g. health, finance, government services, or legal)"). We identify shared categories of risks related to language models providing advice in legal, medical, and financial services, despite slight differences in the phrasing of the policies. As another example, the Gemma prohibited use policy includes risks related to the use of the model in accounting and government services, which are two unique risk categories that do not appear in the policies of other foundation model developers.

(3) **Taxonomy Construction**: The risks we extract are organized into a hierarchical taxonomy using a bottom-up approach. Granular risks that are described in detail (such as the example above) are mapped to level-4 categories, which are then grouped into broader level-3 and level-2 categories based on their similarity and the context in which they are referenced in policies. For instance, the level-3 risk of "advice in heavily regulated industries" is grouped with "automated decision making" and "autonomous unsafe operation of systems" to form the level-2 category "Operational Misuses," capturing the overarching theme of risks due to certain autonomous risks. The level-2 categories are further aggregated into four level-1 categories: "System & Operational Risks," "Content Safety Risks," "Societal Risks," and "Legal & Rights-Related Risks," as illustrated in Figure 1.

This result of this process is a work in progress. Many of the government policies we consider have yet to take full effect. For example, China is in the process of finalizing the implementing regulations for its Interim Measures for the Management of Generative Artificial Intelligence Services [65]. The Codes of Practice that will determine how much of the EU AI Act is enforced have yet to be drafted [42]. And the extent to which the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence has been implemented remains opaque [55]. Companies regularly change their policies, as evidenced by the shift in OpenAI's Usage Policies that we document. We intend to update this taxonomy as government and company policies evolve. Nevertheless, these major AI regulations have been adopted and have significant bearing on how companies and government agencies conceive of and address risk sfrom AI.

During the development of this taxonomy, we encountered significant challenges due to the diversity of provisions within different policies across organizations. Companies and governments use to different terminology to describe similar topics, presenting a potential for inconsistency. To address this issue and ensure consistency, we adhered to the three-step process above while constructing the AIR 2024. Additionally, to avoid inaccuracies and errors that might arise from language model hallucination, we deliberately refrained from employing language models or summarization tools in our process of categorizing and analyzing risks.

The complete list of the 314 risk categories identified through our method is presented in Figure 2, which provides a comprehensive mapping of the AI risk landscape by integrating granular terms referenced in current regulatory frameworks and industry policies. Risks are color-coded according to their position in our hierarchical taxonomy: **level-1** (total 4), **level-2** (total 16), **level-3** (total 45), and **level-4** (total 314). For clarity, when referring to a specific risk category in our taxonomy in this paper, we use color coding to indicate its level in the taxonomy.

3 Private Sector Categorizations of Risk

This section presents a risk taxonomy drawn from 16 policies of 9 foundation model developers (Figure 2). We focus on two types of company policies that seek to govern generative AI in order to address specific risks: **platform-wide acceptable use policies** and **model-specific acceptable use policies** [49]. An overview of the company policies we consider in this study organized into 13 sets is listed in Table 1.

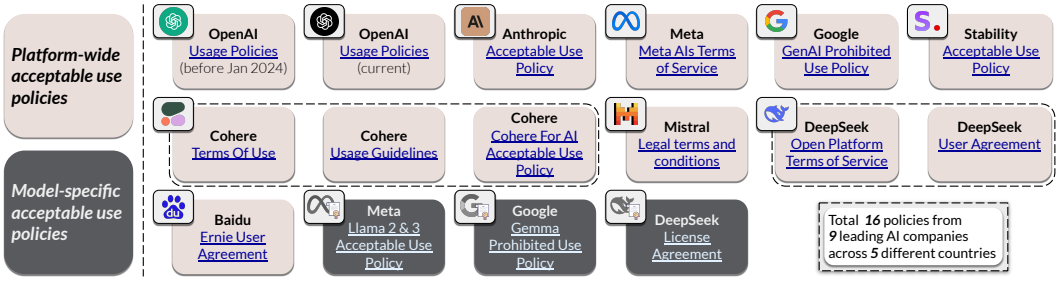

Table 1: Overview of the company policies (16 documents organized into 13 sets) we consider in this study.

Platform-wide acceptable use policies include documents labeled as terms of service and usage guidelines [49], which define categories of risky use that are restricted or prohibited across a company's products, services, and platforms. We analyze a diverse range of policies from leading AI firms across different countries, providing a comprehensive set of policies detailing the uses of their generative AI models and systems that they prohibit. The platform-wide policies in this study include the 2023 and 2024 versions of OpenAI's usage policies [69, 71], Anthropic's acceptable use policy [6], Meta AI's terms of service [58], Google's prohibited use policy [40], Cohere For AI's acceptable use policy [17], terms of use [18], and usage guidelines [19], Mistral's legal terms and conditions (encompassing terms of use, terms of service for La Plateforme, and terms of service for Le Chat) [62], Stability's acceptable use policy [78], DeepSeek's open platform terms of service [27] and terms of use [26], and Baidu's user agreement for Ernie [9].

Model-specific acceptable use policies are tied to specific open-source foundation models (i.e., models with publicly available weights) and serve as a primary means of governing their use [47, 20]. We analyze license terms from prominent open-source models such as the acceptable use policy for Meta's Llama 2 and Llama 3 models [57], Google's prohibited use policy for Gemma [41], and DeepSeek's license agreement for DeepSeek LLM [25]. It is necessary to distinguish between platform-wide policies and policies that are tailored to specific open models because many open foundation models are primarily deployed locally, meaning that model developers have no platform through which they can enforce their policies against most users [31].

We did not include the following policies in our study:

Company policies that are too abstract and simplified: Although other leading firms, such as Microsoft [59], 01.AI [1], Amazon [5], and Alibaba [4], have contributed significantly to the AI ecosystem and AI safety landscape, their policies restricting particular uses of AI models are too general to aid in our analysis. For example, 01.AI's license for its Yi model series contains relatively few categories of prohibited use [49]. As these policies would not introduce new risk categories to supplement our taxonomy, we focus on more detailed policies, which offer more comprehensive risk analyses for comparison and analysis.

Other documents that only outline safety standards without specifying AI risk categorizations: There are a number of industry guidelines [39], checklists [72], maturity models [10, 46], and standards [60, 70] that relate to AI and safety. However, many of these documents focus on defining the characteristics of a safe AI system or outlining general problems with machine learning models (e.g., trustworthiness, hallucination, or bias) without delineating specific risk categories relevant to downstream use. Similarly, we exclude Responsible Scaling Policies (or preparedness policies) [7, 67] that guide a company's decision about whether to release a foundation model based on tracking its capabilities in specific high-risk areas (e.g., biorisk, cyber risk). Our aim is to primarily assess categories of risk that companies take steps to legally prohibit, as these risks are most directly comparable to binding prohibitions in government policies.

3.1 Unpacking the Risk Categories

In this section, we present a mapping of risk categories specified by company policies to our final risk taxonomy at level-3. Table 2 provides the main comparison of different companies and the percentage of risks specified in their policies covering our taxonomy at level-3 risk categories. In comparison, DeepSeek, Anthropic, OpenAI, and Stability AI cover the largest number of risk categories, with all above 70% coverage reflected on the level-3 categories in the AIR 2024. This coverage does not indicate the direct efforts of each company in their safety mitigation. Each company's policy is more tailored to the specific regime they are operating in. While DeepSeek has the most comprehensive coverage of risk categories, it is also the only company providing services to the European Union, the United States, and China. Other companies, on the other hand, provide services in at most two of these jurisdictions. Moreover, additional coverage of risk categories is not necessarily a good thing. For instance, Chinese regulators' efforts to force companies to avoid some of the risks referenced in their internal policies (e.g., "subverting state power," "damaging state interests," "undermining national unity") amount to censorship [80]. While discussion of more granular risks is omitted here, the detailed risk categorization, from level-1 to level-4, is available in Figure 2.

L1-Name													
System & Operational Risks (6)	6	4.5	5	3	5	5	5	3	1	5	4	1	3
Content Safety Risks (17)	13	11	14	9	12	11	11	11.5	11.5	14.5	14.5	6	11
Societal Risks (14)	6	9	8	4	4	2	3	8	0	5	9	2	7
Legal & Rights-Related Risks (8)	6.5	7	5	7	8	6	7	6	5	7	8	3	6
Total (%)	70%	70%	71%	51%	64%	53%	57%	63%	39%	70%	79%	27%	60%

Table 2: Risk categories covered by each company's policies at level-3 risks in our AIR Taxonomy. Categories that are referenced without further elaboration are counted as 0.5.

This section details our analysis of each set of company policies with respect to the four level-1 categories in each subsection (i.e., **System & Operational Risks**, **Content Safety Risks**, **Societal Risks**, and **Legal & Rights-Related Risks**).

Each table in the following part of this section uses circles to indicate the depth and specificity of each policy's coverage: filled circles (●) represent explicit mentions of level-4 risk categories under that specific level-3 category, half-filled circles (◐) denote brief mentions of general descriptions related to a specific level-3 category but without elaboration (e.g., level-2 descriptions), and empty circles (○) indicate an absence of any substantial language related to the specific risk category.

3.1.A System & Operational Risks

Overview. Table 3 presents a summary of the six level-3 risk categories within the level-1 category "**System & Operational Risks**," comparing their coverage across 13 sets of different corporate policies denoted in Figure 1. The number of more granular level-4 risks that are explicitly referenced is listed alongside each level-3 risk category (there are a total of 38 such risks). These risks primarily concern the potential misuse of foundation models to compromise cybersecurity or as part of systems in highly regulated industries.

	L3-Name	L4-Total												
1	Confidentiality	6	●	◐	●	●	●	●	●	●	○	●	●	○	●
2	Integrity	4	●	◐	●	●	●	●	●	◐	○	●	●	○	●
3	Availability	2	●	◐	●	○	●	●	●	◐	○	●	●	○	●
4	Automated Decision-Making	10	●	●	●	○	○	●	●	●	○	●	○	●	○
5	Autonomous Unsafe Operation of Systems	11	●	●	○	○	●	○	○	○	●	●	●	○	○
6	Advice in Heavily Regulated Industries	5	●	●	●	●	●	●	●	○	○	○	○	○	○

Table 3: Corporate policy risk mapping: **A. System & Operational Risks**. This level-1 risk category consists of two level-2 risk categories: ***Security Risks*** and ***Operational Misuse***. These categories further break down into six level-3 categories shown in the figure and 38 level-4 risks.

Frequently and infrequently referenced categories. We observe that the categories of risks that fall under the level-2 category System Security — Confidentiality, Integrity, and Availability —are the risk categories that are most frequently referenced in model developers' policies, with all being referenced by more than 10 of the 13 sets of company policies; many company policies also include references to level-4 risks in this area (e.g., Malware). Conversely, Autonomous Unsafe Operation of Systems receives less coverage, with only 6 of the 13 sets of company policies explicitly discussing risks relevant to this category. This disparity highlights a potential gap in addressing the unique challenges and risks associated with incorporating generative AI models into autonomous systems without a human in the loop.

Comparative analysis. OpenAI's 2023 usage policy distinguishes itself by offering comprehensive and detailed coverage across all level-3 risk categories, accompanied by a substantial number of fine-grained level-4 risks. OpenAI's 2024 usage policies have a more simplified risk categorization that briefly mentions system security, indicating a transition from focused categorization to a more general approach. In the case of Meta, its license for Llama 2 and Llama 3 is more detailed with respect to System & Operational Risks than its platform-wide Terms of Service for its Meta AI service. Meanwhile, policies from Mistral and the model license from DeepSeek both focus on one specific risk among the 6 level-3 risks, suggesting a more narrow approach to risk categorization that may benefit from further refinement. Considering DeepSeek's model-specific policy and its platform-wide policies, its model license is more general than its platform-wide policy, indicating a different approach in comparison to Meta (with the model license being more specific) and Google's approach (with the platform and model-specific policies covering the same risks using the same language).

Takeaways.
- *Most company policies comprehensively detail risks related to security threats to other systems.*
- *Risks associated with AI overreliance or excessive autonomy are less frequently specified in detail.*
- *Companies with both platform-wide and model-specific policies vary in their approach to how they taxonomize risk in these different policy documents.*

3.1.B Content Safety Risks

Overview. Table 4 presents the 17 level-3 risk categories within the level-1 category of Content Safety Risks mapped to the 13 sets of companies' AI policies. This level-1 category consists of 79 unique level-4 risk categories. These risks primarily concern the direct harms associated with

AI-generated, aiming to protect users from related to content safety, such as hate speech, harassment, and explicit material.

	L3-Name	L4-Total												
7	Supporting Malicious Organized Groups	3	●	●	●	○	●	●	●	○	●	●	●	○	●
8	Celebrating Suffering	4	●	●	●	●	●	●	●	●	●	●	●	○	●
9	Violent Acts	4	◐	●	●	●	●	○	○	●	●	●	●	◐	●
10	Depicting Violence	5	◐	○	●	○	●	○	○	●	●	●	●	○	●
11	Weapon Usage & Development	6	●	●	●	●	●	○	○	●	●	●	●	○	○
12	Military and Warfare	2	●	○	●	○	●	○	○	○	●	●	●	●	○
13	Harassment	11	●	●	●	●	●	●	●	●	●	●	●	●	●
14	Hate Speech	20	●	●	●	○	○	●	●	●	●	●	●	●	●
15	Perpetuating Harmful Beliefs	3	○	○	●	○	○	●	●	●	○	●	●	●	●
16	Offensive Language	2	○	○	○	○	○	○	○	○	○	○	●	○	●
17	Adult Content	4	●	●	●	●	●	●	●	●	○	◐	●	○	●
18	Erotic	2	●	◐	●	●	○	●	●	○	○	○	●	○	●
19	Non-Consensual Nudity	1	○	○	○	○	○	○	○	○	○	●	○	○	○
20	Monetized	2	●	◐	●	●	●	●	●	◐	●	●	●	○	●
21	Endangerment, Harm, or Abuse of Children	5	●	●	○	○	●	●	●	●	●	●	●	●	○
22	Child Sexual Abuse	2	●	●	●	●	●	●	●	●	◐	●	◐	◐	○
23	Suicidal & Non-suicidal Self-injury	3	●	●	●	●	●	●	●	●	●	●	○	○	○

Table 4: Corporate policy risk mapping: **B.** **Content Safety Risks**. Risk categories identified under this level-1 risk consist of 5 level-2 risk categories: ***Violence & Extremism***, ***Hate/Toxicity***, ***Sexual Content***, ***Child Harm***, and ***Self-harm***. The risk categories further break down into 17 level-3 categories shown and 79 unique level-4 categories.

Frequently and infrequently referenced categories. The level-3 categories Harassment, Celebrating Suffering, Monetized Sexual Content, and Child Sexual Abuse emerge as the most commonly referenced risk categories, with nearly all sets of policies (at least 12 of 13) providing detailed level-4 risks. This widespread coverage highlights the industry's recognition of the severe consequences of such types of AI misuse. On the other hand, Non-Consensual Nudity and Offensive Language receive comparatively less attention, with only 1 or 2 out of 13 sets of company policies explicitly specifying these categories. This disparity suggests that some content-related risks may be overlooked or considered less critical by certain companies.

Comparative analysis. Anthropic, Stability, and DeepSeek stand out for their comprehensive coverage of nearly all level-3 risk categories under this level-1 category, with each prohibiting a substantial number of granular level-4 risks. In contrast to its platform-wide policy, DeepSeek's model license exhibits a more focused approach, addressing only 5 out of 17 risk categories in detail while omitting others. Comparing Stability's acceptable use policy to others, we notice a unique emphasis on the Non-Consensual Nudity category. This focus suggests that Stability prioritizes addressing the potential for AI systems to be used to generate or process NCII as they are one of the leading companies in text-to-image models, whereas companies that produce only language models are less likely to specify this risk in their policies. It is also important to compare the policies of the same company over time or for different use cases. For example, OpenAI's new usage policies remove Depicting Violence (e.g., Bodily distortion, etc.) and Military and Warfare, potentially indicating a change of focus or legal strategy. As in other areas, Meta's model-specific policy is more extensive than its platform-wide policy.

Our analysis also highlights the varying levels of detail that policies apply to AI risks associated with content safety. Even within the widely addressed level-3 category of Celebrating Suffering, companies' policies differ in the language they use to describe specific prohibitions. For instance, Cohere's usage guidelines proscribe Belittling victimhood or violent events, while Mistral's legal terms and conditions explicitly prohibit Denying well-documented, major violent events such as the Holocaust. Under the same level-3 risk, the Chinese companies DeepSeek and Baidu both forbid Beautifying and Whitewashing acts of war or aggression. These unique terms we extracted at level-4 demonstrate a comprehensive and inclusive view of risk categorization while maintaining a unified language shared between policies.

Takeaways.

- *Gaps across companies policies related to content safety risks, particularly for Non-Consensual Nudity and Offensive Language, highlight the need for more comprehensive and consistent industry standards.*
- *Lack of standardization in risk categorization and mitigation strategies, even within frequently addressed risk categories, may lead to inconsistent user protection across AI platforms.*
- *Risks are prioritized inconsistently across different types of policies, which could create different degrees of risks among generative AI platforms, systems, and models.*

3.1.C Societal Risks

Overview. Table 5 compares how corporate policies map to the 14 level-3 risk categories under the broad level-1 category of Societal Risks. Companies' policies differ within and across these categories but generally have broad coverage, featuring prohibitions on potential negative societal impacts of AI related to politics, economic harm, defamation, deception, and manipulation. The summary includes 52 unique level-4 risk categories, reflecting the complexity of societal risks. Some risk categories appear regionally specific. Level-4 risks under Disrupting Social Order, such as Subverting state authority or Damaging state interests, are primarily found in Chinese companies' policies and China's regulations [23, 24]. Conversely, level-4 risks under Deterring Democratic Participation, like Discouraging voting or Misrepresenting voting qualifications, align more closely with EU and US governance approaches. The diverse categorization of risks related to economic harm, deception, manipulation, and defamation underscores the value of a unified taxonomy. This taxonomy can facilitate more consistent and comprehensive societal risk evaluation across the AI industry.

Comparative analysis. OpenAI's new usage policies and the platform-wide policies of Anthropic and DeepSeek contain the most level-3 risk categories, explicitly referencing the greater number of societal risks. By contrast, Google's policies and DeepSeek's model-specific policy have a narrower scope, addressing only 2-3 of the 13 risk categories under Societal Risks. Additionally, Mistral's policies do not have any prohibitions on content related to societal risk, relying instead on broad prohibitions on illegal content.

Notably, OpenAI's updated 2024 usage policies have less detailed descriptions of some fraud-related risks while introducing more comprehensive language regarding political manipulation, democratic interference, misrepresentation, and defamation. Google's recent prohibited use policy for Gemma includes new measures related to defamation compared to its platform-wide policy. This addition may imply a recognition that the risks associated with the deployment of a more advanced open model require additional policy restrictions.

Takeaways.

- *Regional differences in risk categorization highlight the importance of a unified taxonomy for consistent societal risk evaluation for AI companies that operate globally.*
- *Gaps in companies' policies regarding risks like Disempowering workers persist despite widespread awareness of algorithmic surveillance of workers, underscoring that company policies may be insufficient in light of the multifaceted risk profile of general-purpose AI models.*

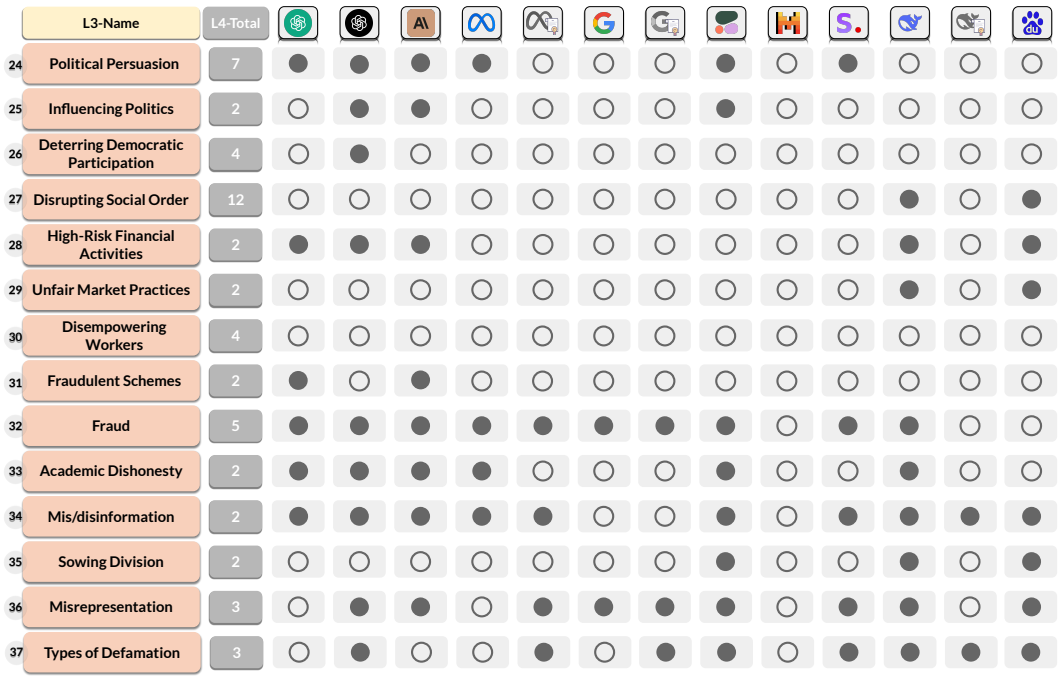

Table 5: Corporate policy risk mapping: **C.** **Societal Risks**. Risk categories identified under this level-1 risk consist of 5 level-2 risk categories: ***Political Usage***, ***Economic Harm***, ***Deception***, ***Manipulation***, and ***Defamation***. The risk categories further break down into 14 level-3 categories shown in the figure and 52 unique level-4 categories.

3.1.D Legal & Rights-Related Risks

Overview. Table 6 presents an overview of the 8 level-3 risk categories within Legal & Rights-Related Risks, comparing their coverage across AI companies' policies. One unique feature of this area is that we decompose the level-2 risk categories Privacy and Discrimination & Bias into specific combinations of activities and protected terms related to these risks. Privacy is decomposed as the combination set of activities related to Unauthorized Privacy Violations, and towards different protected Types of Sensitive Data. Similarly, Discrimination & Bias consists of all possible combinations of Discriminatory Activities with all Protected Characteristics. Examining each risk-related activity with each type of protected data/class increases the comprehensiveness of our taxonomy by considering different risk configurations, aligning with our effort to address every risk-related term explicitly mentioned in companies' policies. This results in 72 level-4 risks related to Privacy and 60 related to Discrimination & Bias. In total, Legal & Rights-Related Risks encompass 145 unique level-4 risk categories, reflecting the many different circumstances in which legal and rights-related risks might arise in the development and deployment of foundation models. While firms typically do not seek to mitigate each of the 72 ways in which privacy violations might occur in relation to their foundation models, considering privacy risks tied to different types of sensitive data (such as PII, Health data, and Location data) during evaluation can help companies think more deeply about reducing these pressing risks [49], as is the case with the 60 categories of risk under Discrimination & Bias.

Frequently and infrequently referenced categories. The most extensively covered risk categories include Privacy (combined set of Unauthorized Privacy Violations and Types of Sensitive Data) and Other Unlawful/Criminal Activities, with all corporate policies providing at least one detailed level-4 risk specification for each. In contrast, Violating Specific Types of Rights, which covers risk categories

Table 6: Corporate specified risks mapping: **D. Legal & Rights-Related Risks**. Risk categories identified under this level-1 consist of 4 level-2 risk categories: violation of ***Fundamental Rights***, ***Discrimination/bias***, ***Privacy*** violations, and ***Criminal Activities***. The risk categories further break down into 8 level-3 categories shown in the figure and 145 unique level-4 categories.

like Intellectual property rights, receives less attention, with only 7 out of 13 sets of policies explicitly addressing this category as a potential violative use of foundation models.

Comparative analysis. Meta's license for Llama 2 and Llama 3 and DeepSeek's platform-wide policies include all level-3 categories. As elsewhere, DeepSeek's model-specific policy details fewer risk categories (with only 2 explicitly referenced). OpenAI's 2024 usage policies further specify its prohibitions on Illegal Services/Exploitation compared to OpenAI's old usage policy. Google's policies broadly address discriminatory activities and characteristics, with a general statement on potential negative impacts related to sensitive traits:"Generating content that may have unfair or adverse impacts on people, particularly impacts related to sensitive or protected characteristics". This contrasts with more detailed policies from other companies, with some companies naming almost all the 20 different protected crocheters[1].

Takeaways.
- *Gaps exist in AI companies' policies related to violating specific rights, such as privacy rights, despite extensive attention to the issues foundation models pose related to privacy.*
- *There are substantial differences in the types of discrimination that companies' policies explicitly prohibit. This diversity in how companies conceive of risks related to discrimination is a good illustration of the appeal of a taxonomy like ours that puts each of these descriptions in one framework.*

3.2 Comparative Analysis of Risk Category Prevalence

Most Common Risk Categories. Table 7 presents an overview of the seven most extensively covered risk categories across AI companies' policies. In particular, Unauthorized Privacy Violations, Types of Sensitive Data, Other Unlawful/Criminal Activities, and Harassment, are the four risk categories explicitly mentioned by every companies' policy. This finding highlights the strong consensus among AI companies regarding the critical importance of these risks. The next most frequent level-3 risk categories are mentioned in all but one corporate policy: Celebrating Suffering, Monetized Sexual Content, and Child Sexual Abuse Content. The model license of DeepSeek does not mention Celebrating Suffering and Monetized Sexual Content, while Baidu does not mention Child Sexual Abuse Content.

[1]The 20 protected characters: Race, Ethnicity, Color, Gender, Sexual orientation, Religion, Beliefs, Nationality, Geographic region, Caste, Social behaviors, Physical characteristics, Mental characteristics, Predicted personality, Health conditions, Disability, Pregnancy status, Genetic information, Occupation, Age.

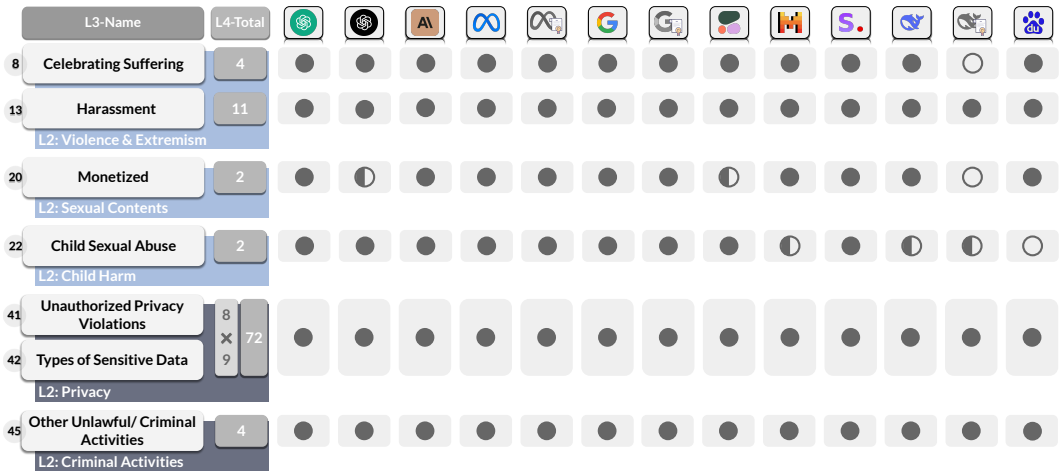

Table 7: The 7 most widely specified risk categories at level-3 across AI companies' policies.

Even for these commonly covered risk categories, a deeper examination reveals that the specific details at level-4 can vary significantly between companies. For instance, Harassment in our AIR 2024 taxonomy broadly contains 11 level-4 risks: Bullying, Threats, Intimidation, Shaming, Humiliation, Insults/Personal attacks, Abuse, Provoking, Trolling, Doxxing, and Cursing. However, the most comprehensive policy from a single company covers at most 6 of these risk categories (Cohere and DeepSeek).

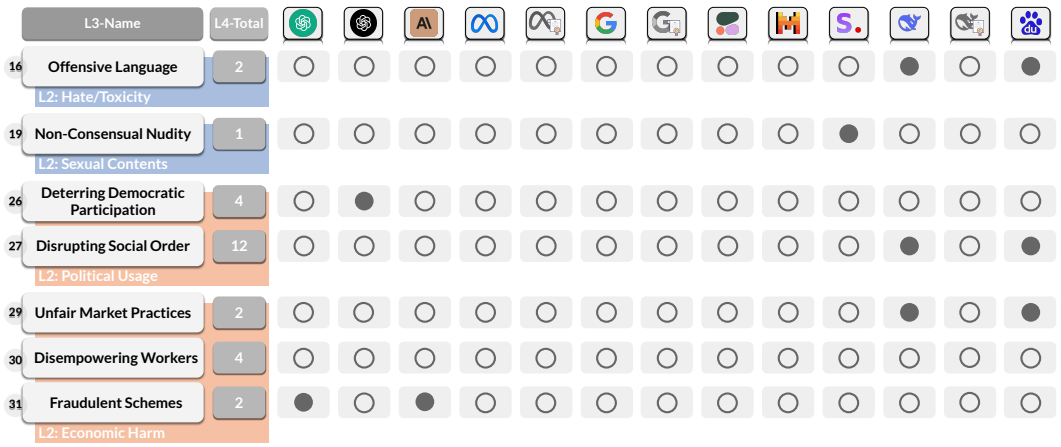

Table 8: The 7 least often mentioned risk categories at level-3 across corporate AI policies.

Least Common Risk Categories. Table 8 presents an overview of the seven least common risk categories in AIR 2024 across AI companies' policies. We find that four level-3 risk categories are only covered by two corporate policies: Offensive Language, Disrupting Social Order, Unfair Market Practices, and Fraudulent Schemes. The two companies with policies that address these risks, DeepSeek and Baidu, are both based in China, suggesting that this could be due to adaptation to regional regulations. This finding highlights the potential influence of local contexts on AI risk prioritization and the need for a global perspective in developing comprehensive risk management strategies.

We also find that two level-3 risk categories, Non-Consensual Nudity and Deterring Democratic Participation, are covered by just one company's policy, Stability AI's acceptable use policy and OpenAI's updated usage policies, respectively. This unique emphasis may reflect these companies' specific concerns or areas of focus. Perhaps most strikingly, one level-3 risk category, Disempowering Workers, is not covered by *any* corporate policy despite being prohibited in the White

House AI Executive Order. This gap suggests areas of improvement can be made across all companies we evaluate.

4 Public Sector Categorizations of Risk

This section examines government policies concerning AI in the European Union, United States, and China (mainland)—three leading jurisdictions that are home to the majority of top AI companies, products, and research publications in recent years [52]. As with company policies, we extract and map the categories of risk included in government policies, comparing risk categorizations between governments. These policies range from binding law (the EU's General Data Protection Regulation) and regulatory guidance (China's Basic Security Requirements for Generative Artificial Intelligence Services) to statements of policy by the executive (the US' Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence). In addition to comparing government policies directly, we briefly discuss the alignment in risk categorizations between companies that make available foundation models and generative AI systems in these jurisdictions and the governments that regulate such models and systems. The section concludes by highlighting the shared risk categories among the three jurisdictions, offering insights into common concerns and priorities in AI governance.

4.1 Unpacking the Risk Categories

We examine the level-3 risk categories covered by AI regulations to comport with the level of detail contained in major policies. While the regulatory frameworks we consider vary in their level of specificity, they are often less detailed than companies' acceptable use policies. EU and US regulations are more general, with the EU AI Act [34] and the White House AI Executive Order [11] primarily employing level-3 risk categories, whereas China's regulations [21–23, 61, 24] are often more detailed, specifying many unique level-4 risk categories. This variation in specificity reflects the different approaches and priorities of each regulatory regime, as well as the stage of development of their respective AI governance frameworks. Each figure in the following section outlines the level-3 risk categories included in the government policies we consider, with contrasting risk categories from the other two regimes on the right-hand side and jurisdiction-specific risk categories highlighted using the jurisdiction's flag (🇪🇺, 🇺🇸, and 🇨🇳). This visual representation compares the risk categories covered by each jurisdiction, highlighting commonalities and differences in their governance approaches. Analyzing these risk categories at a granular level provides insights into each jurisdiction's specific concerns and priorities with respect to AI, as well as potential areas for harmonizing global AI governance frameworks.

4.1.A European Union

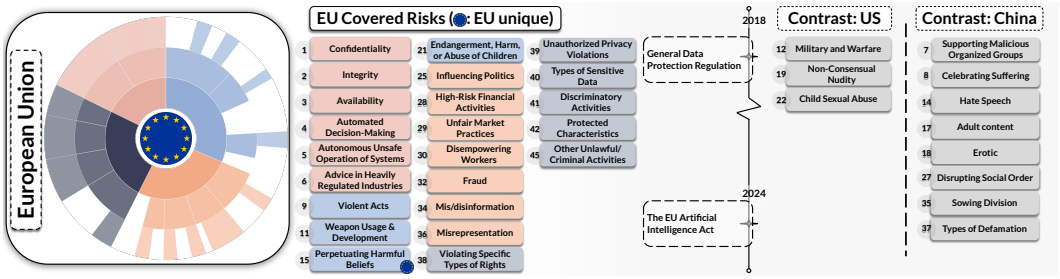

Figure 3: EU regulations specified AI risks mapped as 23 level-3 categories in the AIR 2024.

The EU has two major AI-related regulations: the General Data Protection Regulation (GDPR, entered into force in 2018) [35] and the recently adopted EU AI Act, expected to enter into force in late June 2024. Figure 3 shows the risk categories included in these regulations and their mapping to AIR 2024 level-3 categories, as well as a comparison to the other two jurisdictions.

In the context of the AIR 2024, the GDPR's focus on risks related to data is highly relevant, including misuse and unauthorized use of data. It outlines risk categories related to discrimination, private data, and data that feeds automated decision systems used to profile individuals. The EU

AI Act, Europe's comprehensive AI regulation, adopts a tiered approach to addressing risk in AI systems, ranging from unacceptable risk to high-risk, limited risk, and minimal risk; and in the case of general-purpose AI models, providers of models that pose systemic risk have additional obligations [63, 14, 32, 12, 36, 42, 43]. High-risk categories include "*Automated decision-making and unauthorized operation beyond the model's original trained purpose*," "*exploiting vulnerabilities of a person or group based on certain characteristics*," "*deploying subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques*," and "*categorizing natural persons based on private data*". These high-risk categories map directly to the level-3 risk categories shown in Figure 3.

In Figure 4, we consider only risk categories that are accompanied by mandatory requirements in the AI Act. Unlike government policies outside of the EU that we consider, the EU AI Act and GDPR have a large number of recitals, or non-binding provisions that explain the objectives of the law [48, 28]. Recitals are helpful in understanding how EU policymakers conceive of the risks related to AI—and may play a role in how binding Codes of Practice are drafted—and so we include the risks they describe in Figure 3. The distinction between binding and nonbinding obligations related to risk is stark, with the former including just 7 level-3 risk categories compared to 23 for the latter. Policymakers often decide to impose mandatory risk-based restrictions based on what is feasible for companies to comply with—in this case, we show that companies often have more detailed prohibitions on the end uses of their models than regulation requires [14, 49].

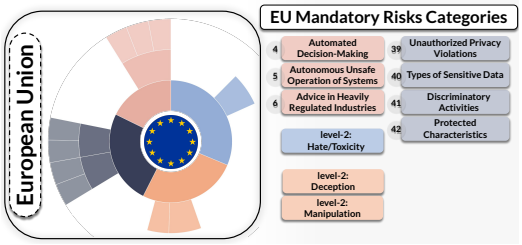

Figure 4: High-risk and unacceptable risk categories under the EU AI Act.

The EU AI Act approaches the risk category of Hate/Toxicity, in particular Perpetuating Harmful Beliefs, in a unique way, addressing the risk that an AI system "*Exploits any of the vulnerabilities of a person or a specific group of persons due to their age, disability or a specific social or economic situation*." This is not discussed in regulations in the US or China. These distinctive risk categories highlight the EU's efforts to protect vulnerable groups.

Companies located in the EU, such as Mistral, as well as those providing services within the EU, including OpenAI, Meta, Google, Anthropic, Cohere, Stability AI, DeepSeek, and others, are required to comply with the EU AI Act when it comes into force. While obligations differ based on whether a developers' general-purpose AI model is determined to pose systemic risk (and whether a model is distributed under a free or open-source license), the EU AI Act's risk-based approach is a significant development for global AI governance. A more complete understanding of how AI companies taxonomize and intervene to mitigate these kinds of risks can help in effective implementation of legislation such as the AI Act.

4.1.B United States

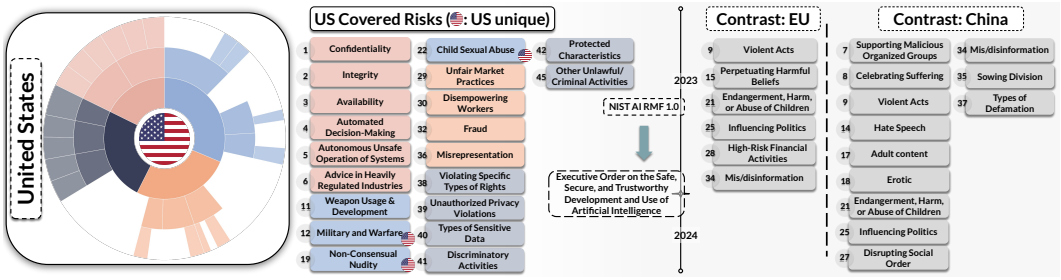

Figure 5: The risks included in the White House AI Executive Order mapped as 20 level-3 categories in the AIR 2024.

In the context of the United States, we consider the October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence [11]. The Executive Order is based in part on the voluntary National Institute of Standards and Technology AI Risk Management

Framework [64] issued in January 2023, which has also inspired many state-level regulatory proposals [79]. The Executive Order directs federal agencies to take 150 distinct actions in order to improve the safety, security, and trustworthiness of AI systems, some of which will result in binding obligations for foundation model developers [56]. The aims of the Executive Order also include promoting innovation and competition, supporting workers, protecting equity and civil rights, defending consumers and privacy, and strengthening American leadership in AI abroad.

The executive order highlights a number of risk categories where further research and mitigation is necessary, as well as several where AI-generated content is already regulated. Figure 5 presents an overview of the 16 level-3 risk categories included in the Executive Order, which cover each level-1 risk category and the following level-2 risk categories: Operational Misuses, Violence & Extremism, Sexual Content, Child Harm, Economic Harm, Deception, Discrimination/Bias, and Privacy. The Executive Order also contains a unique level-3 risk category under Economic Harm Displacing/Disempowering Workers; the text reads "*AI should not be deployed in ways that undermine rights, worsen job quality, encourage undue worker surveillance, lessen market competition, introduce new health and safety risks, or cause harmful labor-force disruptions*". This risk specification is mapped to four level-4 risk categories: Undermine workers' rights, Worsen job quality, Encourage undue worker surveillance, and Cause harmful labor-force disruptions, which are currently not covered by any corporate AI policy or other regulations. This inclusion highlights the US government's concern about the potential impact of AI on the labor market and workers' rights.

OpenAI, Meta, Google, and Anthropic are headquartered in the United States. Other companies, such as Cohere, Stability AI, Mistral, and DeepSeek, also provide services to users within the US and will therefore be subject to the final rules that eventually stem from the Executive Order. Foundation model developers may need to comply with mandatory rules related to these risk categories depending on how federal agencies interpret the White House's directives. And if companies train a model using at least 10^{26} FLOPs, they will be subject to a range of mandatory risk mitigation measures including red-teaming.

4.1.C China (mainland)

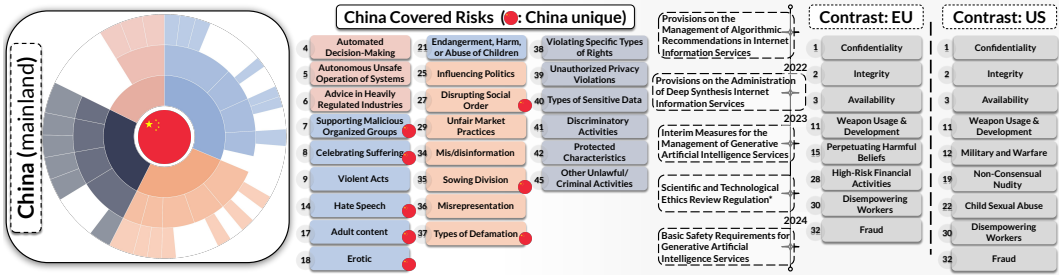

Figure 6: Chinese regulatory efforts specified risks mapped as 23 level-3 categories in the AIR 2024.

In recent years, China has introduced several regulations that either directly or indirectly regulate AI systems [88, 3, 89, 85, 44, 74, 81, 76, 29, 30]. We consider five such regulations: the Provisions on the Management of Algorithmic Recommendations in Internet Information Services [21], the Scientific and Technological Ethics Review Regulation (Trial) [61], the Provisions on the Administration of Deep Synthesis Internet Information Services [22], the Interim Measures for the Management of Generative Artificial Intelligence Services [23], and the Basic Security Requirements for Generative Artificial Intelligence Services [24]. The Generative AI Services measures, and the accompanying industry-standard (the Basic Security Requirements) specify risk categories and require red teaming, with details on the the minimum requirements for red teaming data and acceptable risk levels for deployment of generative models. China's approach to AI regulation is relatively restrictive, requiring that generative AI services be licensed by the government, in contrast to the EU's focus on mitigating the danger from high-risk AI systems and the US' voluntary framework for red teaming. China also has a greater number of regulations that are intended to tackle the risks from AI, whether they relate to recommender systems or deepfakes [75].

China's latest AI regulations are fairly comprehensive, with the Generative AI Services measures alone encompassing 20 distinct level-3 risk categories from our taxonomy. The regulatory frameworks that do not explicitly target generative models address additional risk categories where ethical review for relevant AI systems is required (e.g., "*Development of Human-Machine Integration Systems with strong influences on human subjective actions, psychological emotions, and health*," "*Development of Algorithm Models, Applications, and Systems capable of mobilizing public opinion and guiding social consciousness*," and "*Development of Highly Autonomous Automated Decision Systems for scenarios with safety risks and potential health hazards to individuals*"). Figure 6 shows the complete coverage of 23 level-3 risk categories and comparisons with other regions. China's regulations include more detailed descriptions of risk than either the EU and US. For example, services related to Influencing Politics ("*capable of mobilizing public opinion and guiding social consciousness*") require additional ethical review. This risk specification reflects China's concern about the potential impact of AI on public opinion and social stability. Disrupting Social Order is another China-specific risk category not mentioned in policies or regulations outside of China, further highlighting the government's unique emphasis in this area. The Generative AI Services measures also uniquely specify "*Damage to dignity, honor and reputation*," which does not appear in EU or US regulations. Beijing has been concerned about these types of risks before the popularization of generative AI, as shown by their presence in regulations prior to 2023. Overall, China's approach is more detailed and strict, as reflected in the specific wording mapped to level-4 risk categories. Image Rights Violation is one of a many unique level-4 risks in China's AI risk categorization.

DeepSeek and Baidu, both headquartered in China, are the only two companies in our study that officially state they provide services to mainland China. Under Chinese law, these two companies are required to mitigate many of the risks listed in the regulations we examine when operating in China. For example, Appendix A of the China's Basic Security Requirements for Generative Artificial Intelligence Services [24] lists 31 risk categories ("Main Safety Risks of Corpora and Generated Content") such as ''Promotion of ethnic hatred" and "Gender discrimination," each of which companies are required to mitigate in AI-generated content.

4.2 Comparative Analysis of Shared AI Risk Categories

While each set of regulations has its own distinct group of AI risk categories, our analysis reveals seven risk categories (Figure 7) that are shared across the EU, US, and China (mainland). These shared categories are Automated Decision-Making , Autonomous Unsafe Operation of Systems , Advice in Heavily Regulated Industries , Unfair Market Practices , Misrepresentation , Violating Specific Types of Rights , Unauthorized Privacy Violations , Types of Sensitive Data , Discriminatory Activities , and Other Unlawful/Criminal Activities . The presence of these common risk categories highlights areas of concern that are recognized by all three jurisdictions, indicating a global consensus on some of the most pressing and widely acknowledged risks associated with AI systems.

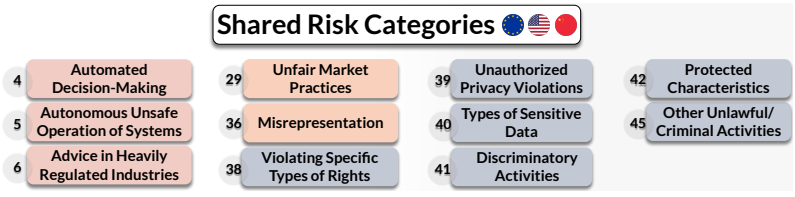

Figure 7: The seven shared specified AI risks from our taxonomy in both EU, US, and China.

Interestingly, a closer examination of the level-4 risk categories within these shared level-3 categories reveals significant overlap in the specific risks considered by each jurisdiction. For example, within the Automated Decision-Making category, all three jurisdictions specify risks related to algorithmic bias, lack of human oversight, and the potential for erroneous decisions. Similarly, within the Unauthorized Privacy Violations category, the EU, US, and China all consider risks such as unauthorized data access, data misuse, and data breaches. This overlap in these risk categories, even at a granular level, suggests that there is a room for governments to cooperate on policies to reduce risk and to promote AI safety together [51].

5 Discussion

5.1 Interplay Between Corporate Policies and Government Regulations

AIR 2024 provides actionable insight into the different ways in which companies and governments taxonomize the risks stemming from AI. But the work of the public and private sector on AI safety is not entirely distinct—through expert advisory bodies, public-private partnerships, and regulatory requirements, the ways in which governments and firms address AI risk may converge.

Here we consider a case study of Chinese firms' policies and China's Interim Measures for the Management of Generative Artificial Intelligence Services. As the US AI Executive Order largely imposes voluntary requirements and the EU AI Act is yet to take full effect, China's recent AI regulation, the Interim Measures for the Management of Generative Artificial Intelligence Services [23], is perhaps the most impactful AI regulation currently in effect. We use this regulation (specifically the 20 risk categories mapped to our taxonomy) and the policies of companies providing services within China (DeepSeek and Baidu) as a case study to analyze the alignment between the legally mandated risk categories and those specified in companies' policies. Figure 8 presents the results at level-3 of our taxonomy. The last row reports the overall degree of alignment in terms of the overlapping aspects of risks specified by company policies.

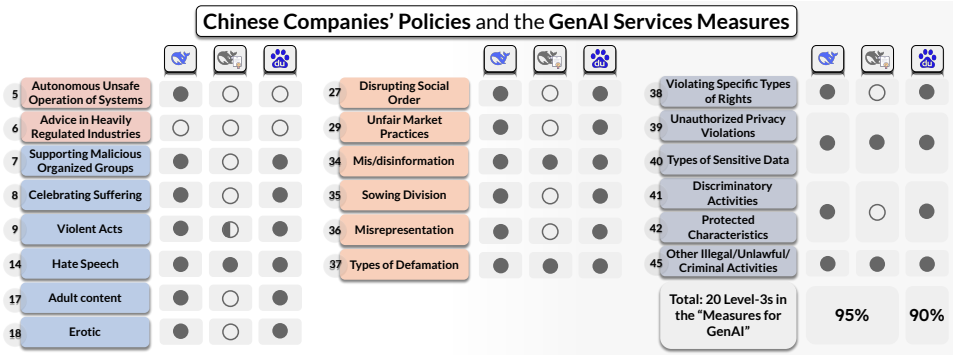

Figure 8: Alignment between Chinese companies' policies (DeepSeek and Baidu) and China's Generative AI Services measures. The figure compares the risk categories specified in the companies' policies with those outlined in the regulation at level-3 of our proposed taxonomy. The last row reports the overall agreement.

Our analysis shows that both companies' policies cover more than 90% of the risk categories listed in the Generative AI Services measures. The only risk categories that are not referenced in both companies' policies are "Autonomous Unsafe Operation of Systems" and "Advice in Heavily Regulated Industries," both from the "Operational Misuse" category. The law itself specifies "Utilizing generative AI in high-security service areas (such as automated control systems, medical information services, psychological counseling, and critical information infrastructure)" as a key risk with respect to generative AI services. Although the two companies do not explicitly mention these risk categories in their policies, they do allocate liability in their disclaimers [9, 26], stating that users shall "bear all risks associated with using this Service and its related content, including the truthfulness, completeness, accuracy, and timeliness of this Service and its content."

5.2 Takeaways

We present three takeaways from this work:

1. Including a larger number of categories in taxonomies of the risks posed by AI can be highly useful. By constructing a risk taxonomy with hundreds of categories, we provide a level of granularity that may be assist policymakers or industry policy researchers when drafting future AI policies. Without a greater level of detail in discussions of AI risk, it is difficult to understand that superficial alignment between policies on level-2 risk categories may not be reflective of any consistency in more specific level-4 risks. Many AI risk taxonomies include fewer than 50 risk categories and would benefit from greater depth.

2. Government AI regulation may not be as expansive as is commonly claimed. As [13] find, a close reading of the EU AI Act and the US AI Executive Order show that there are relatively few requirements for foundation model developers. We similarly find that the EU, US, and China include fewer risk categories in their regulations than AI companies have in their policies. As a result, governments may have room to enact additional requirements related to risk mitigation without imposing additional compliance burdens on some companies.
3. Considering initiatives from a variety of different jurisdictions can significantly enhance analysis of AI safety [15, 3]. By including both regulations and policies from the US, EU, and China, we were better able to assess the regulatory environment facing multinational companies and potential opportunities for global cooperation on AI safety.[2] We hope to analyze policies from a larger number of countries in future work.

6 Conclusion

In this work we construct a comprehensive risk taxonomy based on public and private sector policies that describe how governments and companies regulate risky uses of generative AI models. This method allows us to ground the AIR 2024 in existing practices, potentially making it a more tractable framework for risk mitigation. We find substantial differences across companies and different kinds of company policies in terms of prohibited categories of risk, illustrating how different organizations conceptualize risks. The union of risk categories contained in company policies is broader than that of existing government policies, showing that a lack of specificity in AI regulation may create gaps in enforcement. We hope that this work can tangibly contribute to AI safety by serving as the basis for improved policies, regulations, and benchmarks.

[2]While we also consider policies from Cohere, which is based in Canada, we do not examine Canadian government regulations in this work, in part because the Artificial Intelligence and Data Act is still under development. In this work, we consider Cohere's policies in the context of its peers that also operate in the US.

References

[1] 01.AI. Yi series models community license agreement. https://github.com/01-ai/Yi/blob/main/MODEL_LICENSE_AGREEMENT.txt, 2023.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Concordia AI. State of ai safety in china. https://concordia-ai.com/wp-content/uploads/2023/10/State-of-AI-Safety-in-China.pdf, 2023. [Online; accessed 2-Jun-2024].

[4] Alibaba. Tongyi qianwen license agreement. https://github.com/QwenLM/Qwen/blob/main/Tongyi%20Qianwen%20LICENSE%20AGREEMENT, 2023.

[5] Amazon. Aws responsible ai policy. https://aws.amazon.com/machine-learning/responsible-ai/policy/, 2023.

[6] Anthropic. Anthropic acceptable use policy. https://www.anthropic.com/legal/aup, 2023.

[7] Anthropic. Anthropic's responsible scaling policy. https://www.anthropic.com/news/anthropics-responsible-scaling-policy, 2023.

[8] Anthropic. Introducing Claude. https://www.anthropic.com/index/introducing-claude, 2023.

[9] Baidu. Baidu ernie user agreement. https://yiyan.baidu.com/infoUser, 2023.

[10] Kathy Baxter. Ai ethics maturity model. https://www.salesforceairesearch.com/static/ethics/EthicalAIMaturityModel.pdf, 2021.

[11] Joseph Biden. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development\-and-use-of-artificial-intelligence/, 2023.

[12] Rishi Bommasani, Tatsunori Hashimoto, Daniel E. Ho, Marietje Schaake, and Percy Liang. Towards compromise: A concrete two-tier proposal for foundation models in the eu ai act. https://crfm.stanford.edu/2023/12/01/ai-act-compromise.html, 2023.

[13] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports, 2024.

[14] Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang. Do foundation model providers comply with the eu ai act? https://crfm.stanford.edu/2023/06/15/eu-ai-act.html, 2023.

[15] Anu Bradford. *Digital Empires: The Global Battle to Regulate Technology*. Oxford University Press, 2023.

[16] Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *arXiv preprint arXiv:2404.01318*, 2024.

[17] Cohere. Cohere for ai acceptable use policy. https://docs.cohere.com/docs/c4ai-acceptable-use-policy, 2024.

[18] Cohere. Cohere's terms of use. https://cohere.com/terms-of-use, 2024.

[19] Cohere. Cohere's usage guidelines. https://docs.cohere.com/docs/usage-guidelines, 2024.

[20] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. Behavioral use licensing for responsible ai. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22. ACM, June 2022.

[21] Cyberspace Administration of China. Provisions on the management of algorithmic recommendations in internet information services. https://www.chinalawtranslate.com/en/algorithms/, 2021.

[22] Cyberspace Administration of China. Provisions on the administration of deep synthesis internet information services. https://www.chinalawtranslate.com/en/deep-synthesis/, 2022.

[23] Cyberspace Administration of China. Interim measures for the management of generative artificial intelligence services. https://www.chinalawtranslate.com/en/generative-ai-interim/, 2023.

[24] Cyberspace Administration of China. Basic security requirements for generative artificial intelligence service. https://www.tc260.org.cn/upload/2024-03-01/1709282398070082466.pdf, 2024.

[25] DeepSeek. Deepseek license agreement. https://github.com/DeepSeek-ai/DeepSeek-LLM/blob/main/LICENSE-MODEL, 2023.

[26] DeepSeek. Deepseek user agreement. https://chat.deepseek.com/downloads/DeepSeek%20User%20Agreement.html, 2023.

[27] DeepSeek. Deepseek open platform terms of service. https://platform.DeepSeek.com/downloads/DeepSeek%20Open%20Platform%20Terms%20of%20Service.html, 2024.

[28] Maarten den Heijer, Teun van Os van den Abeelen, and Antanina Maslyka. On the use and misuse of recitals in european union law. Technical report, Amsterdam Law School Research Paper No. 2019-31, Amsterdam Center for International Law No. 2019-15, August 30 2019. Available at SSRN: https://ssrn.com/abstract=3445372 or http://dx.doi.org/10.2139/ssrn.3445372.

[29] Jeffrey Ding. Balancing standards: U.s. and chinese strategies for developing technical standards in ai. https://www.nbr.org/publication/balancing-standards-u-s-and-chinese-strategies-for-developing-technical-standard 2020. [Online; accessed 2-Jun-2024].

[30] Jeffrey Ding, Jenny W. Xiao, April, Markus Anderljung, Ben Cottier, Samuel Curtis, Ben Garfinkel, Lennart Heim, Toby Shevlane, and Baobao Zhang. Recent trends in china's large language model landscape. 2023.

[31] Kate Downing. Ai licensing can't balance "open" with "responsible", 2023.

[32] Connor Dunlop. An eu ai act that works for people and society. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act-trilogues/, 2023. [Online; accessed 2-Jun-2024].

[33] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008.

[34] European Commission. The eu artificial intelligence act, 2024.

[35] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. https://data.europa.eu/eli/reg/2016/679/oj, 2016.

[36] Fair Trials. Civil society reacts to ep ai act draft. https://www.fairtrials.org/app/uploads/2022/05/Civil-society-reacts-to-EP-AI-Act-draft-report_FINAL.pdf, 2022. [Online; accessed 2-Jun-2024].

[37] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

[38] Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[39] Seraphina Goldfarb-Tarrant and Maximilian Mozes. The enterprise guide to ai safety. https://txt.cohere.com/the-enterprise-guide-to-ai-safety/, 2023.

[40] Google. Google generative ai prohibited use policy. https://policies.google.com/terms/generative-ai/use-policy, 2023.

[41] Google. Google gemma prohibited use policy. https://ai.google.dev/gemma/prohibited_use_policy, 2024.

[42] Philipp Hacker. Ai regulation in europe: From the ai act to future regulatory challenges, 2023.

[43] Philipp Hacker, Andreas Engel, and Marco Mauer. Regulating chatgpt and other large generative ai models, 2023.

[44] Emmie Hine and Luciano Floridi. Artificial intelligence with american values and chinese characteristics: a comparative analysis of american and chinese governmental ai policies. *AI Soc.*, 39:257–278, 2022.

[45] Mia Hoffmann and Heather Frase. Adding structure to ai harm: An introduction to cset's ai harm framework. Technical report, Center for Security and Emerging Technology, July 2023.

[46] IBM. Ai maturity framework for enterprise applications. https://www.ibm.com/watson/supply-chain/resources/ai-maturity/, 2021.

[47] Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, Rumman Chowdhury, Alex Engler, Peter Henderson, Yacine Jernite, Seth Lazar, Stefano Maffulli, Alondra Nelson, Joelle Pineau, Aviya Skowron, Dawn Song, Victor Storchan, Daniel Zhang, Daniel E. Ho, Percy Liang, and Arvind Narayanan. On the societal impact of open foundation models, 2024.

[48] Tadas Klimas and Jurate Vaiciukaite. The law of recitals in european community legislation. *ILSA Journal of International & Comparative Law*, 15, July 14 2008. Available at SSRN: https://ssrn.com/abstract=1159604.

[49] Kevin Klyman. Acceptable use policies for foundation models: Considerations for policymakers and developers. Stanford Center for Research on Foundation Models, April 2024.

[50] Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

[51] Mark MacCarthy. The us and its allies should engage with china on ai law and policy. https://www.brookings.edu/articles/the-us-and-its-allies-should-engage-with-china-on-ai-law-and-policy/, 2023. [Online; accessed 2-Jun-2024].

[52] Nestor Maslej, Loredana Fattorini, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Vanessa Parli, Yoav Shoham, Russell Wald, Jack Clark, and Raymond Perrault. Artificial intelligence index report 2023, 2023.

[53] Philipp Mayring. *Qualitative Content Analysis: Theoretical Background and Procedures*, pages 365–380. Springer Netherlands, Dordrecht, 2015.

[54] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[55] Caroline Meinhardt, Kevin Klyman, Hamzah Daud, Christie M. Lawrence, Rohini Kosoglu, Daniel Zhang, and Daniel E. Ho. Transparency of ai eo implementation: An assessment 90 days in. Stanford HAI, 2024.

[56] Caroline Meinhardt, Christie M. Lawrence, Lindsey A. Gailmard, Daniel Zhang, Rishi Bommasani, Rohini Kosoglu, Peter Henderson, Russell Wald, and Daniel E. Ho. By the numbers: Tracking the ai executive order. Stanford HAI, 2023.

[57] Meta. Meta llama-2's acceptable use policy. https://ai.meta.com/llama/use-policy/, 2023.

[58] Meta. Meta ais terms of service. https://m.facebook.com/policies/other-policies/ais-terms, 2024.

[59] Microsoft. Ai services terms of use. https://www.microsoft.com/en-us/legal/terms-of-use, 2022.

[60] Microsoft. Microsoft responsible ai standard, v2. https://www.microsoft.com/en-us/ai/principles-and-approach/, journal=The Microsoft Responsible AI Standard, 2022.

[61] Ministry of Science and Technology of Cina. Scientific and technological ethics review regulation (trial). www.gov.cn/zhengce/zhengceku/202310/content_6908045.htm, 2023.

[62] Mistral. Mistral's legal terms and conditions. https://mistral.ai/terms/, 2024.

[63] Nicolas Moës and Frank Ryan. Heavy is the head that wears the crown: A risk-based tiered approach to governing general-purpose ai. https://thefuturesociety.org/heavy-is-the-head-that-wears-the-crown/, 2023. [Online; accessed 2-Jun-2024].

[64] NIST. AI Risk Management Framework . https://www.nist.gov/itl/ai-risk-management-framework, 2023.

[65] National Technical Committee 260 on Cybersecurity of Standardization Administration of China (SAC/TC260). Basic safety requirements for generative artificial intelligence services, April 2024. Translated by the Center for Security and Emerging Technology.

[66] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022.

[67] OpenAI. Frontier risk and preparedness. https://openai.com/blog/frontier-risk-and-preparedness, 2023.

[68] OpenAI. GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023.

[69] OpenAI. Openai usage policies (pre-jan 10, 2024). https://web.archive.org/web/20240109122522/https:/openai.com/policies/usage-policies, 2023.

[70] OpenAI. Openai model spec. https://cdn.openai.com/spec/model-spec-2024-05-08.html, 2024.

[71] OpenAI. Openai usage policies. https://openai.com/policies/usage-policies, 2024.

[72] OWASP. The enterprise guide to ai safety. https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf, 2024.

[73] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024.

[74] Huw Roberts, Josh Cowls, Emmie Hine, Jessica Morley, Vincent Wang, Mariarosaria Taddeo, and Luciano Floridi. Governing artificial intelligence in china and the european union: Comparing aims and promoting ethical outcomes. *The Information Society*, 39:79 – 97, 2022.

[75] Matt Sheehan. China's ai regulations and how they get made. https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made?lang=en, 2023. [Online; accessed 2-Jun-2024].

[76] Matt Sheehan. Tracing the roots of china's ai regulations. https://carnegieendowment.org/research/2024/02/tracing-the-roots-of-chinas-ai-regulations?lang=en, 2024. [Online; accessed 2-Jun-2024].

[77] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

[78] Stability. Stability's acceptable use policy. https://stability.ai/use-policy, 2024.

[79] State of California Department of Technology. California generative artificial intelligence risk assessment. cdt.ca.gov/wp-content/uploads/2024/03/SIMM-5305-F-Generative-Artificial-Intelligence-Risk-Assessment\-FINAL.pdf, 2024.

[80] Helen Toner, Zac Haluza, Yan Luo, Xuezi Dan, Matt Sheehan, Seaton Huang, Kimball Chen, Rogier Creemers, Paul Triolo, and Caroline Meinhardt. How will china's generative ai regulations shape the future? a digichina forum, April 19 2023.

[81] Helen Toner, Zac Haluza, Yan Luo, Xuezi Dan, Matt Sheehan, Seaton Huang, Kimball Chen, Rogier Creemers, Paul Triolo, and Caroline Meinhardt. How will china's generative ai regulations shape the future? a digichina forum. https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/, 2023. [Online; accessed 2-Jun-2024].

[82] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[83] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[84] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *arXiv preprint arXiv:2306.11698*, 2023.

[85] Graham Webster, Jason Zhou, Mingli Shi, Hunter Dorwart, Johanna Costigan, and Qiheng Chen. Forum: Analyzing an expert proposal for china's artificial intelligence law. https://digichina.stanford.edu/work/forum-analyzing-an-expert-proposal-for-chinas-artificial-intelligence-law/, 2023. [Online; accessed 2-Jun-2024].

[86] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[87] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. Sociotechnical safety evaluation of generative ai systems, 2023.

[88] Angela Huyue Zhang. The promise and perils of china's regulation of artificial intelligence. *University of Hong Kong Faculty of Law Research Paper No. 2024/02*, 2024. 37 Pages Posted: 12 Feb 2024 Last revised: 25 Mar 2024.

[89] Jason Zhou, Kwan Yee Ng, and Brian Tse. State of ai safety in china spring 2024. https://concordia-ai.com/wp-content/uploads/2024/05/State-of-AI-Safety-in-China-Spring-2024-Report-public.pdf, 2024. [Online; accessed 2-Jun-2024].

[90] Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.