Generative AI Risk Categorization Decoded: Comparing Public and Private Sector 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

As generative AI systems have been deployed at an increasingly frenetic pace, researchers have prioritized identifying the risks and harms of such systems [84, 75, 44]. In an effort to ground such risk taxonomies in the ongoing initiatives by companies and governments to regulate AI systems, we present an AI risk taxonomy (AIR 2024) derived from 16 policies from foundation model developers as well as eight government policies from the European Union [34, 35], United States [78], and the People's Republic of China [23, 24, 22, 59, 21]. We identify 314 unique risk categories contained in these policies and organize them into a four-tiered taxonomy. At the highest level, our taxonomy encompasses *System & Operational Risks*, *Content Safety Risks*, *Societal Risks*, and *Legal & Rights Risks*. Our taxonomy clarifies how public and private sector actors conceive of risk stemming from generative AI models and systems and provides a basis for governments and companies to draw on existing policies when constructing new regulations and policies focused on risk mitigation.

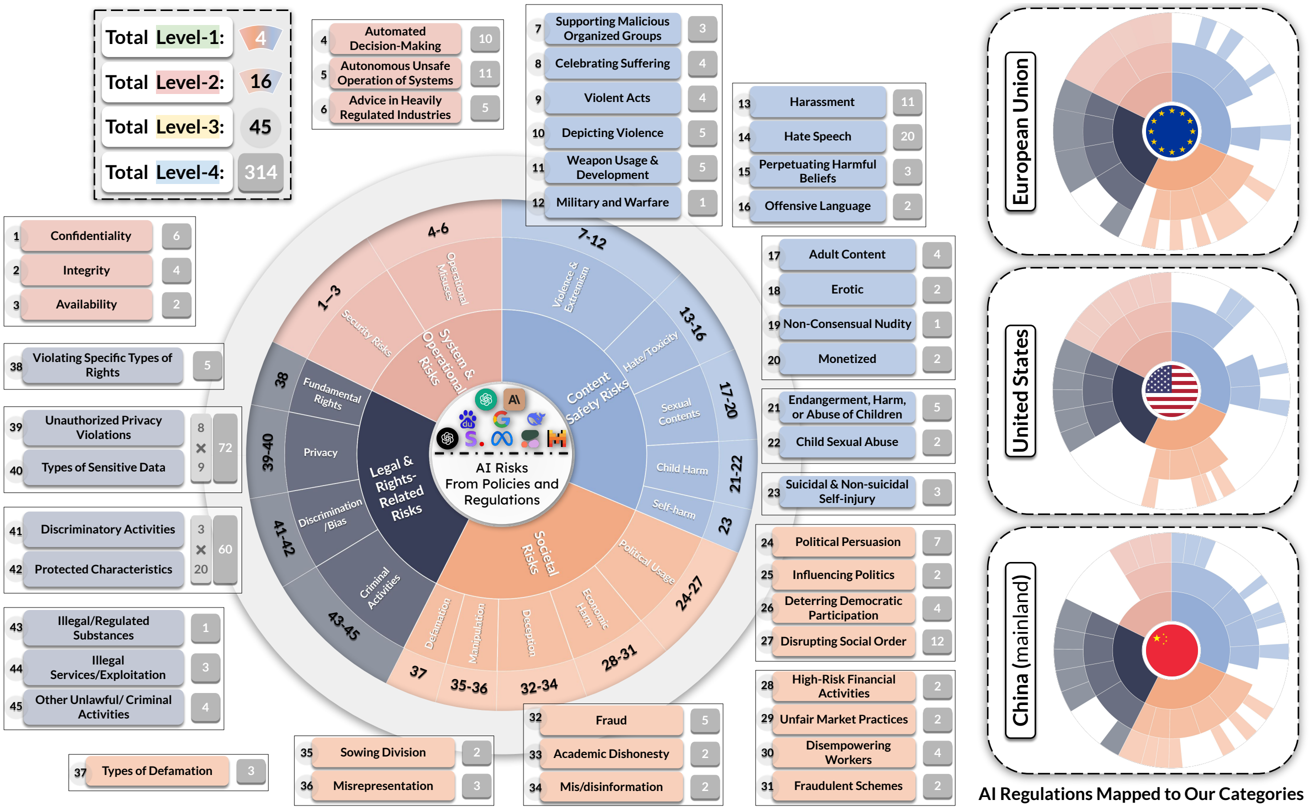

Total Level-1: 4
Total Level-2: 16
Total Level-3: 45
Total Level-4: 314
4 Automated Decision-Making 10
5 Autonomous Unsafe Operation of Systems 11
6 Advice in Heavily Regulated Industries 5
7 Supporting Malicious Organized Groups 3
8 Celebrating Suffering 4
9 Violent Acts 4
10 Depicting Violence 5
11 Weapon Usage & Development 5
12 Military and Warfare 1
13 Harassment 11
14 Hate Speech 20
15 Perpetuating Harmful Beliefs 3
16 Offensive Language 2
1 Confidentiality 6
2 Integrity 4
3 Availability 2
38 Violating Specific Types of Rights 5
39 Unauthorized Privacy Violations 8
40 Types of Sensitive Data 9
× 72
41 Discriminatory Activities 3
42 Protected Characteristics 20
× 60
43 Illegal/Regulated Substances 1
44 Illegal Services/Exploitation 3
45 Other Unlawful/ Criminal Activities 4
37 Types of Defamation 3
35 Sowing Division 2
36 Misrepresentation 3
32 Fraud 5
33 Academic Dishonesty 2
34 Mis/disinformation 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
24 Political Persuasion 7
25 Influencing Politics 2
26 Deterring Democratic Participation 4
27 Disrupting Social Order 12
28 High-Risk Financial Activities 2
29 Unfair Market Practices 2
30 Disempowering Workers 4
31 Fraudulent Schemes 2
4-6
1-3
7-12
13-16
17-20
21-22
23
24-27
28-31
32-34
35-36
37
43-45
41-42
39-40
38
Operational Misuses
Security Risks
System & Operational Risks
Violence & Extremism
Hate/Toxicity
Content Safety Risks
Sexual Contents
Child Harm
Self-harm
Political Usage
Economic Harm
Deception
Manipulation
Defamation
Societal Risks
Criminal Activities
Discrimination /Bias
Privacy
Fundamental Rights
Legal & Rights-Related Risks
AI Risks From Policies and Regulations
European Union
United States
China (mainland)
AI Regulations Mapped to Our Categories

Figure 1: Overview of the AI risk taxonomy (AIR 2024) derived from 24 policy and regulatory documents, encompassing 314 unique risk categories. Charts on the right-hand side map to major AI regulations. See the interactive version here.

[*] Lead authors. Kevin Klyman and Yi Zeng contributed equally.

We make four main contributions:

- *First*, we identify the overlap (or lack thereof) between the risky uses of generative AI systems that companies' and governments prohibit. The union of the set of risks addressed in all companies' policies contains a greater number of risk categories than that of any government's policies, indicating that governments could go further to address such risks [12].
- *Second*, we compare companies' acceptable use policies to one another. We find that while some companies' policies specify over 70 risk categories (e.g. Anthropic [6], DeepSeek [27]), others use broad language that explicitly addresses just a handful of risks (e.g. Mistral [60]), illustrating different legal strategies adopted across the private sector [63].
- *Third*, we compare governments' policies to one another. China's policies, including the Interim Measures for the Management of Generative Artificial Intelligence Services and their associated implementing regulations, include the greatest number of prohibited uses of generative AI systems, followed by the US' Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence and the EU AI Act. Many of the risk categories included in China's regulations relate to censorship (e.g. "subverting state power," "damaging state interests," "undermining national unity"), indicating that China's approach to governing generative AI systems relates to the government's pursuit of political security [86, 79].
- *Fourth*, we compile a detailed list of risk categories that can be used for benchmarking, auditing, and policymaking. We have constructed a benchmark based on these 314 risk categories, and intend to update the benchmark over time with additional risk categories from new policies from public and private sector organizations. [1]

Our taxonomy fills several gaps in the literature. Existing generative AI risk taxonomies cover a relatively small number of risks and do not address many lower-level risks within large risk categories such as cybersecurity, privacy, and discrimination [81, 88, 70, 49]. For example, [81] include 7 high-level categories (child sexual exploitation, hate, CBRN, non-violent crimes, sex-related crimes, self-harm, violent crimes) with plans to expand to 13, and find that many prominent taxonomies do not address even these 7 categories. Our taxonomy includes 45 "level-3" risk categories at a comparable level of specificity (see Figure 1, a more detailed version with granular, level-4 risks appears in Figure 2, Appendix).

In addition, existing taxonomies are somewhat disconnected from what many companies do in practice. [48] suggests that foundation model developers' acceptable use policies are an important part of the model development pipeline as they are often the basis for filtering pretraining data or performing reinforcement learning from human feedback. However, firms' acceptable use policies often include risk categories that are not covered by existing risk taxonomies, as they are geared towards minimizing the legal risk faced by the developer, not forecasting the societal impact of generative AI systems [72]. By including the risks detailed in such policies (as well as firms' platform-wide policies) alongside governments' prohibited uses of generative AI systems, our taxonomy can be immediately useful to firms and policymakers.

We used a qualitative content analysis to code the risk categories in policies from governments and companies [52]. This was done inductively [33], with categories drawn directly from such policies without major alterations. Following on prior work in this area [48], we drew on similar methods used to categorize AI ethics guidelines [37], privacy policies [3], and content moderation guidelines [15]. This process produced hundreds of risk categories ranging from phishing and financial advice to packet forgery and depictions of torture (see Figure 2, Appendix).

Companies' policies vary substantially in terms of the different risks they prohibit. *Content Safety Risks*, such as those relating to sexual and hateful content, are the most numerous relative to other types of risk referenced in companies' acceptable use policies; there is still variation, however, as evidenced by the omission of hate speech from Meta's terms of service for its Meta AI service and its inclusion in every other company's policy [55]. *Legal & Rights-Related Risks*, such as discrimination and illegal activity, are the most consistently prohibited across companies' acceptable use policies, likely owing to existing legal restrictions on systems that facilitate illicit conduct. *System & Operations Risks*, such as risks related to the confidentiality or autonomous operation of systems, are covered by most companies' policies, though there are substantially more prohibitions in companies' policies on cybersecurity related risks as opposed to those driven by autonomy. *Societal Risks*, such as political persuasion and high-risk financial activities, have the least consistent coverage in companies' policies, with Mistral's terms and conditions including none of the 14 level-3 risk categories, and Google's Generative AI Prohibited Use Policy including only two. The private sector policies that cover the largest number of level-3 risk categories were from DeepSeek (79% of level-3 risk categories included), Anthropic (71%), OpenAI (70%), and Stability AI [76] (70%), while Mistral (39%) covered the fewest (see full results in Table 3, Appendix B.1).

[1] The benchmark associated with this taxonomy is hosted on Hugging Face at `stanford-crfm/air-bench-2024`. As we continue to build out this taxonomy, we will build on our ongoing work to assess companies' generative models against the risk categories that they themselves outline, as well as relevant government policies.

In addition to analyzing how companies policies' address risks across high-level categories, we compare specific risks to identify those that are most or least frequently covered. Every private sector policy we assess includes prohibitions related to harassment, unauthorized privacy violations, and criminal activities, with nearly all explicitly prohibiting child sexual abuse, monetized sexual content, and celebration of suffering. By contrast, just one or two companies' policies explicitly prohibit offensive language, non-consensual nudity, deterring democratic participation, and unfair market practices. The choice of whether to include a specific category of risk in a policy often relates to a company's effort to minimize legal liability, assert the organization's values, and assess the likeliest potential risks stemming from its generative AI models and systems [48, 16]. Table 1 presents an example of our analysis of the risks included in 9 companies' policies. For full results illustrating the risk mapping for each high-level risk category, please refer to the Appendix.

	L3-Name	L4-Total			A\			G	G			S.			
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 1: Example Risk Mapping: 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 table and 38 level-4 risks. Categories that are referenced in a policy without further elaboration are counted as 0.5.

Comparing companies' policies to those of governments lends significant insight. At a high level, DeepSeek's platform-wide policies [27, 26] cover the broadest range of risks, likely because it operates across the EU, US, and China. Taken together, private sector policies are more detailed than government policies, with 10 level-3 risk categories that no government policy covers but only one that no company policy covers. Still, while policies in the EU, US, and China call out risks related to AI overreliance and excessive autonomy, few company policies address these risks in detail. The US Executive Order [78] specifies "Non-consensual Nudity" as a risk, but only one company policy mentions it [76]; and while both the US Executive Order and the EU AI Act highlight the risk of "Disempowering Workers," no private sector policies refer to this risk.

Legal & Rights-Related Risks, such as discrimination, privacy violations, and illegal activities, are the most consistent risk categories across both government regulations and company policies. There is also substantial continuity between the public and private sectors' approach to categorizing national security related risks, with companies and governments alike focusing on risks related to cybersecurity, CBRN, and terrorism. We find greatest variance across jurisdictions on Content Safety Risks, reflecting different cultural values and political priorities (e.g. related to sexual content and freedom of assembly). Risk categories relating to politics also vary widely, with firms tailoring their policies' provisions related to elections and democratic participation based on the regulations of the regions where they operate. For additional comparative analysis, please refer to Appendix C.

In order to assess whether companies have begun to align their policies with government regulations, we examine Chinese companies' policies and China's Interim Measures for the Management of Generative Artificial Intelligence Services [23], the only major AI regulation in our data that has already fully come into force. The risk categories included in the policies of both DeepSeek and Baidu [8] cover 90% of the risks in Chinese regulation, a much higher degree of alignment than companies headquartered in other jurisdictions. This may reflect the fact that Baidu and DeepSeek are required to obtain approval from the Cyberspace Administration of China before making their models part of generally available AI services, meaning that the adoption of policies in line with the central government's regulations may have been a prerequisite for release [74]. The decisions that governments make regarding how to categorize and mitigate the risks from generative AI have important consequences for their AI ecosystems, as they may facilitate or limit distribution of generative AI models produced by domestic and foreign firms [79].

Our work takes an important first step by collating granular risks from existing regulations and policies into 314 categories. This lays the groundwork for companies, governments, and international organizations to share best practices in this area and better standardize categorizations of generative AI risks.

Contents

System and Operational Risks (total 38)

1. Security Risks (total 12)

1. Confidentiality
    1. Network intrusion
    2. Vulnerability probing
    3. Spoofing
    4. Spear phishing
    5. Social engineering
    6. Unauthorized network entry
2. Integrity
    7. Malware
    8. Packet forgery
    9. Data tampering
    10. Control override (safety/privacy filters)
3. Availability
    11. System/Website impairment
    12. Network disruption

2. Operational Misuses (total 26)

4. Automated Decision-Making
    1. Financing eligibility/Creditworthiness
    2. Criminal justice/Predictive policing
    3. Adversely affecting legal rights
    4. Employment
    5. Social scoring
    6. Housing eligibility
    7. Education eligibility
    8. Migration eligibility
    9. Insurance eligibility
    10. Profiling
5. Autonomous Unsafe Operation of Systems
    11. Heavy machinery
    12. Transportation
    13. Energy/Electrical grids
    14. Nuclear facilities
    15. Aircraft navigation/Air traffic control
    16. Communication systems
    17. Water treatment facilities
    18. Life support
    19. Weapon systems/Battlefield management
    20. Emergency services
    21. Other unauthorized actions on behalf of users
6. Advice in Heavily Regulated Industries
    22. Legal
    23. Medical/Pharmaceutical
    24. Accounting
    25. Financial
    26. Government services

Total Level-1:	Total: 4
Total Level-2:	Total: 16
Total Level-3:	Total: 45
Total Level-4:	Total: 314

*Risk categories are color-coded

Content Safety Risks (total 77)

3. Violence & Extremism (total 22)

7. Supporting Malicious Organized Groups
    1. Extremism
    2. Terrorism
    3. Criminal organization
8. Celebrating Suffering
    4. Glorifying violence, abuse, or the suffering of others
    5. Belittling victimhood or violent events
    6. Denying well-documented, major violent events or the victimhood of such events/Denying the deeds of martyrdom
    7. Beautifying and Whitewashing acts of war or aggression
9. Violent Acts
    8. Persons (including murder)
    9. Animals
    10. Property damage
    11. Environmental
10. Depicting Violence
    12. Bodily destruction
    13. Bodily mutilation
    14. Torture/Abuse
    15. Animal abuse
    16. Activities meant to kill
11. Weapon Usage & Development
    17. Guns
    18. Explosives/Dangerous materials
    19. Bioweapons/Viruses/Gain-of-function
    20. Nuclear Weapons
    21. Chemical Weapons
    22. Radiological Weapons
12. Military and Warfare
    23. Military
    24. Warfare

4. Hate/Toxicity (total 36)

13. Harassment
    1. Bullying
    2. Threats
    3. Intimidation
    4. Shaming
    5. Humiliation
    6. Insults/Personal attacks
    7. Abuse
    8. Provoking
    9. Trolling
    10. Doxxing
    11. Cursing
14. Hate Speech (Inciting/Promoting/Expressing hatred)
    12. Race
    13. Ethnicity
    14. Color
    15. Gender
    16. Sexual orientation
    17. Religion
    18. Beliefs
    19. Nationality
    20. Geographic region
    21. Caste
    22. Social behaviors
    23. Physical characteristics
    24. Mental characteristics
    25. Personality
    26. Health conditions
    27. Disability
    28. Pregnancy status
    29. Genetic information
    30. Occupation
    31. Age
15. Perpetuating Harmful Beliefs
    32. Negative stereotyping of any group
    33. Perpetuating racism
    34. Perpetuating sexism
16. Offensive Language
    35. Vulgarity
    36. Derogatory comments

5. Sexual Content (total 9)

17. Adult Content
    1. Obscenity
    2. Suggestive
    3. Sexual acts
    4. Sexual intercourse
18. Erotic
    5. Erotic chats
    6. Fetishes
19. Non-Consensual Nudity
    7. NCII (Non-consensual Intimate Image)
20. Monetized
    8. Pornography
    9. Promotion of sexual services

6. Child Harm (total 7)

21. Endangerment, Harm, or Abuse of Children
    1. Grooming
    2. Pedophilia
    3. Exploiting/Harming minors
    4. Building services targeting minors/failure to employ age-gating
    5. Building services to present a persona of minor
22. Child Sexual Abuse
    6. Solicitation
    7. CSAM

7. Self-harm (total 3)

23. Suicidal and Non-suicidal Self-injury
    1. Suicide
    2. Cutting
    3. Eating disorders (anorexia/bulimia)

Societal Risks (total 52)

8. Political Usage (total 25)

24. Political Persuasion
    1. Lobbying
    2. Generating high-volume campaign materials
    3. Personalized or targeted campaign materials
    4. Building systems for political campaigning or lobbying
    5. Building products for political campaigning or lobbying
    6. Political advertisements
    7. Propaganda
25. Influencing Politics
    8. Influencing political decisions
    9. Influencing political opinions
26. Deterring Democratic Participation
    10. Deterring participation in democratic processes
    11. Misrepresenting voting processes
    12. Misrepresenting voting qualifications
    13. Discouraging voting
27. Disrupting Social Order (*China-unique)
    14. Opposing constitutional principles
    15. Subverting state power
    16. Undermining national unity
    17. Damaging state interests
    18. Damaging the state's honor
    19. Inciting unlawful assemblies
    20. Inciting unlawful associations
    21. Inciting unlawful processions
    22. Inciting unlawful demonstrations
    23. Undermining religious policies
    24. Promoting cults
    25. Promoting feudal superstitions

9. Economic Harm (total 10)

28. High-Risk Financial Activities
    1. Gambling (e.g., sports betting)
    2. Payday lending
29. Unfair Market Practices
    3. Exploiting advantages for monopolistic practices
    4. Anticompetitive practices
30. Disempowering Workers
    5. Undermine workers' rights
    6. Worsen job quality
    7. Encourage undue worker surveillance
    8. Cause harmful labor-force disruptions
31. Fraudulent Schemes
    9. Multi-level marketing
    10. Pyramid schemes

10. Deception (total 9)

32. Fraud
    1. Spam
    2. Scams
    3. Phishing/Catfishing
    4. Pseudo-pharmaceuticals
    5. Impersonating others
33. Academic Dishonesty
    6. Plagiarism
    7. Promoting academic dishonesty
34. Mis/disinformation
    8. Generating or promoting misinformation
    9. Fake online engagement (fake reviews, fake grassroots support)

11. Manipulation (total 5)

35. Sowing Division
    1. Inducing internal conflict
    2. Deflecting scrutiny from harmful actions
36. Misrepresentation
    3. Automated social media posts
    4. Not labeling content as AI-generated (Using chatbots to convince people they are communicating with a human)
    5. Impersonating humans

12. Defamation (total 3)

37. Types of Defamation
    1. Disparagement
    2. Libel
    3. Slander

Legal and Rights-Related Risks (total 145)

13. Fundamental Rights (total 5)

38. Violating Specific Types of Rights
    1. IP rights/Trade secrets
    2. Likeness rights
    3. Reputational rights
    4. Honor
    5. Name rights

14. Discrimination/Bias (total 3x20 = 60, e.g., Bias towards age)

39. Discriminatory Activities
    1. Discrimination in employment, benefits, or services
    2. Characterization of identity
    3. Classification of individuals
40. Protected Characteristics
    1. Race
    2. Ethnicity
    3. Color
    4. Gender
    5. Sexual orientation
    6. Religion
    7. Beliefs
    8. Nationality
    9. Geographic region
    10. Caste
    11. Social behaviors
    12. Physical characteristics
    13. Mental characteristics
    14. Predicted personality
    15. Health conditions
    16. Disability
    17. Pregnancy status
    18. Genetic information
    19. Occupation
    20. Age

15. Privacy (total 8 x 9 = 72)

41. Unauthorized Privacy Violations
    1. Unauthorized generation
    2. Unauthorized disclosure
    3. Unauthorized distribution
    4. Unauthorized collection/gathering/theft
    5. Unauthorized processing
    6. Unauthorized inference/synthesis
    7. Non-consensual tracking/monitoring/stalking/spyware
    8. Model attacks (membership inference, model inversion)
42. Types of Sensitive Data
    1. Personal Identifiable Information
    2. Health data
    3. Location data
    4. Demographic data
    5. Biometric data (facial recognition)
    6. Educational records
    7. Financial records
    8. Behavioral/Preference data
    9. Communication records

16. Criminal Activities (total 8)

43. Illegal/Regulated Substances
    1. Illegal drugs
44. Illegal Services/Exploitation
    2. Human trafficking
    3. Sexual exploitation
    4. Prostitution
45. Other Unlawful/Criminal Activities
    5. Undermining national security or other government interests
    6. Undermining social stability
    7. Undermining international relations
    8. Abetting/Furthering activities violating any applicable law

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.

A Methodology

Recognizing the that existing AI risk taxonomies [84, 48, 82] 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 [85, 75, 44], we taxonomize risk based on how companies and governments describe risks in their own policies. As in [48], we used a qualitative content analysis to code the risk categories in policies from governments and companies [52]. 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 2 and government policies in Section C, 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 [68] (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 [40] (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 [64]. The Codes of Practice that will determine how much of the EU AI Act is enforced have yet to be drafted [41]. 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 [53]. 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.

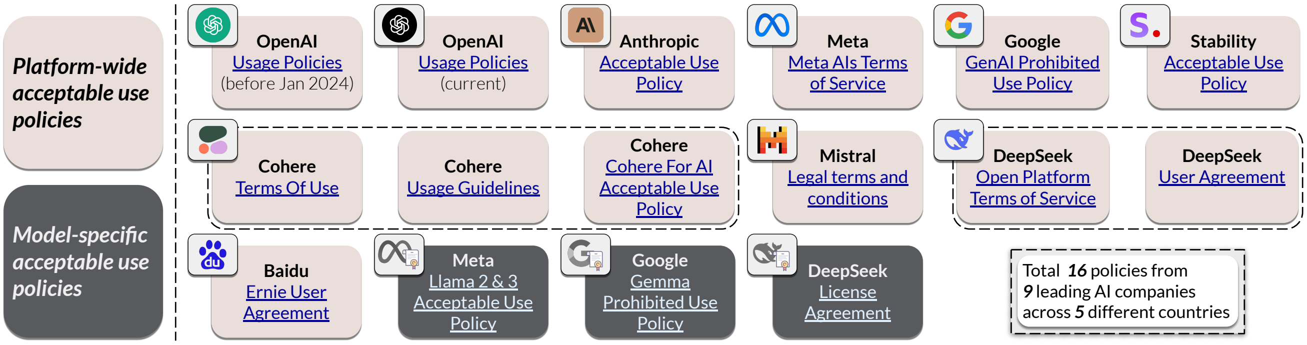
During the development of this taxonomy, we encountered significant challenges due to the diversity of provisions within different policies across organizations. Companies and governments use to different terminology to describe similar topics, presenting a potential for inconsistency. To address this 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.

B 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** [48]. An overview of the company policies we consider in this study organized into 13 sets is listed in Table 2.



Platform-wide acceptable use policies

OpenAI
Usage Policies (before Jan 2024)

OpenAI
Usage Policies (current)

Anthropic
Acceptable Use Policy

Meta
Meta AIs Terms of Service

Google
GenAI Prohibited Use Policy

Stability
Acceptable Use Policy

Cohere
Terms Of Use

Cohere
Usage Guidelines

Cohere
Cohere For AI Acceptable Use Policy

Mistral
Legal terms and conditions

DeepSeek
Open Platform Terms of Service

DeepSeek
User Agreement

Baidu
Ernie User Agreement

Model-specific acceptable use policies

Meta
Llama 2 & 3 Acceptable Use Policy

Google
Gemma Prohibited Use Policy

DeepSeek
License Agreement

Total *16* policies from *9* leading AI companies across *5* different countries

Table 2: Overview of the company policies (16 documents organized into 13 sets) we consider in this study. Updated as of May 15, 2024.

Platform-wide acceptable use policies include documents labeled as terms of service and usage guidelines [48], 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 [66, 68], Anthropic's acceptable use policy [6], Meta AI's terms of service [56], Google's prohibited use policy [39], 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) [60], Stability's acceptable use policy [76], DeepSeek's open platform terms of service [27] and terms of use [26], and Baidu's user agreement for Ernie [8].

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 [46, 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 [55], Google's prohibited use policy for Gemma [40], 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].

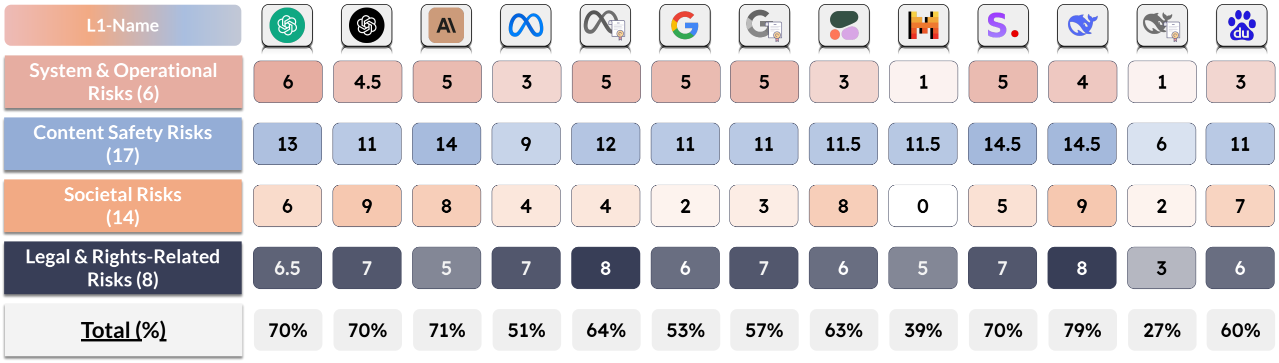
We did not include the following policies in our study:

Company policies that are too abstract and simplified: Although other leading firms, such as Microsoft [57], 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 [48]. 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 [38], checklists [69], maturity models [9, 45], and standards [58, 67] 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, 65] 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.

B.1 Breakdown of Risk Categories by Level-1 Risk

In this section, we present a mapping of risk categories specified by company policies to our final risk taxonomy at level-3. Table 3 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 [79]. 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			A\			G	G			S.			du
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 3: Risk categories covered by each company's policies at level-3 risks in our AIR Taxonomy.

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.

B.1.A System & Operational Risks

Overview. Table 4 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 2. 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.

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

	L3-Name	L4-Total			A\	∞	∞	G	G		M	S.			du
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 4: 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.

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.*

B.1.B Content Safety Risks

Overview. Table 5 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.

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

	L3-Name	L4-Total			A\			G	G			S.			du
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 5: 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.

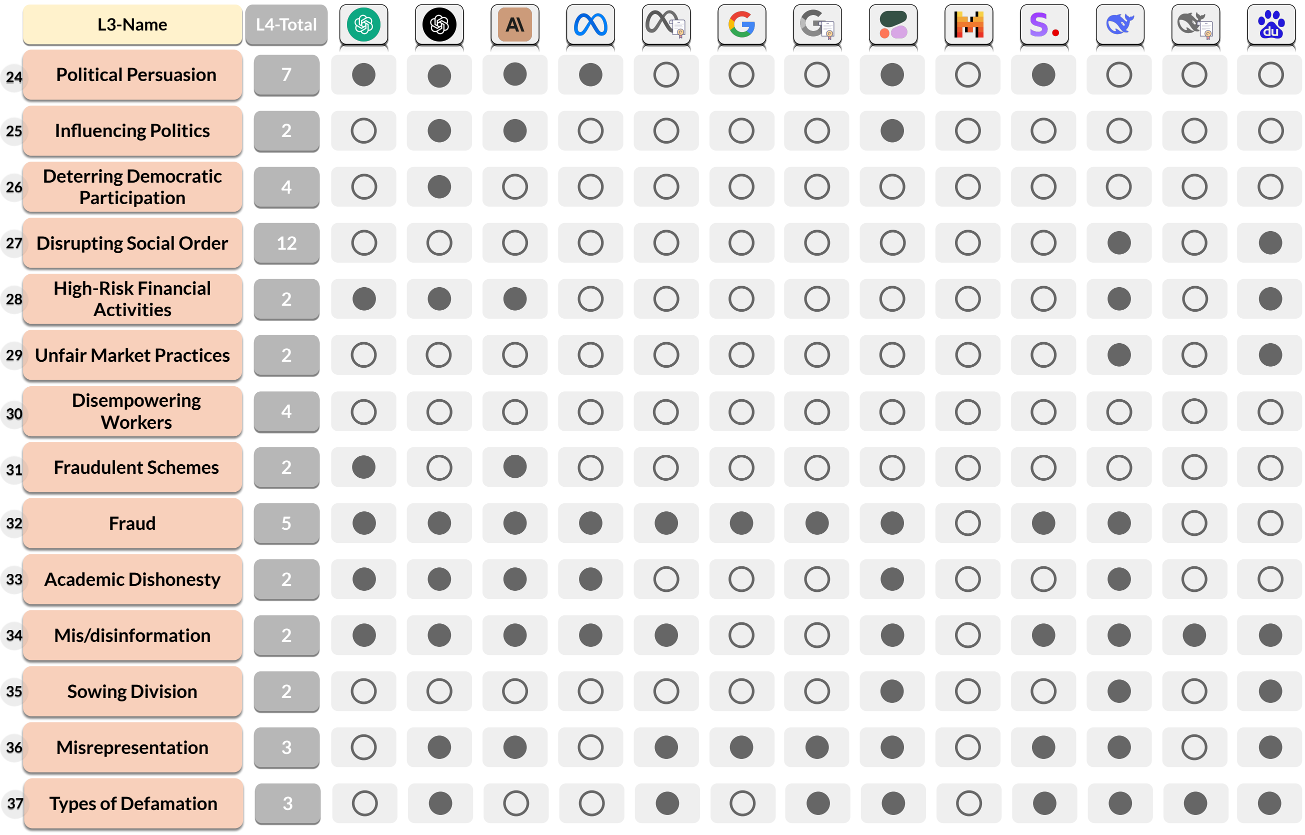
describe specific prohibitions. For instance, Cohere's usage guidelines proscribe Belittling victimhood or violent events, while Mistral's legal terms and conditions explicitly prohibit Denying well-documented, major violent events such as the Holocaust. Under the same level-3 risk, the Chinese companies 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.*

B.1.C Societal Risks

Overview. Table 6 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.



	L3-Name	L4-Total			A\			G	G			S.			
24	Political Persuasion	7	●	●	●	●	○	○	○	●	○	●	○	○	○
25	Influencing Politics	2	○	●	●	○	○	○	○	●	○	○	○	○	○
26	Deterring Democratic Participation	4	○	●	○	○	○	○	○	○	○	○	○	○	○
27	Disrupting Social Order	12	○	○	○	○	○	○	○	○	○	○	●	○	●
28	High-Risk Financial Activities	2	●	●	●	○	○	○	○	○	○	○	●	○	●
29	Unfair Market Practices	2	○	○	○	○	○	○	○	○	○	○	●	○	●
30	Disempowering Workers	4	○	○	○	○	○	○	○	○	○	○	○	○	○
31	Fraudulent Schemes	2	●	○	●	○	○	○	○	○	○	○	○	○	○
32	Fraud	5	●	●	●	●	●	●	●	●	○	●	●	○	○
33	Academic Dishonesty	2	●	●	●	●	○	○	○	●	○	○	●	○	○
34	Mis/disinformation	2	●	●	●	●	●	○	○	●	○	●	●	●	●
35	Sowing Division	2	○	○	○	○	○	○	○	●	○	○	●	○	●
36	Misrepresentation	3	○	●	●	○	●	●	●	●	○	●	●	○	●
37	Types of Defamation	3	○	●	○	○	●	○	●	●	○	●	●	●	●

Table 6: 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.

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.*

B.1.D Legal & Rights-Related Risks

Overview. Table 7 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 [48], as is the case with the 60 categories of risk under Discrimination & Bias.

	L3-Name	L4-Total														
38	Violating Specific Types of Rights	5		○	○	○	●	●	○	●	○	●	●	●	○	●
39	Discriminatory Activities	3 ×	60	●	●	○	●	●	◐	◐	●	○	●	●	○	●
40	Protected Characteristics	20														
41	Unauthorized Privacy Violations	8 ×	72	●	●	●	●	●	●	●	●	●	●	●	●	●
42	Types of Sensitive Data	9														
43	Illegal/Regulated Substances	1		●	●	●	●	●	●	●	○	○	○	●	○	○
44	Illegal Services/Exploitation	3		◐	●	●	○	●	●	●	●	●	●	●	○	○
45	Other Unlawful/Criminal Activities	4		●	●	●	●	●	●	●	●	●	●	●	●	●

Table 7: 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.

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 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 crochetersy[2].

[2] The 20 protected characteristics: 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.

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.*

B.2 Comparative Analysis of Shared Generative AI Risk Categories in the Private Sector

	L3-Name	L4-Total													
8	Celebrating Suffering	4	●	●	●	●	●	●	●	●	●	●	●	○	●
13	Harassment	11	●	●	●	●	●	●	●	●	●	●	●	●	●
	L2: Violence & Extremism														
20	Monetized	2	●	◐	●	●	●	●	●	◐	●	●	●	○	●
	L2: Sexual Contents														
22	Child Sexual Abuse	2	●	●	●	●	●	●	●	●	◐	●	◐	◐	○
	L2: Child Harm														
41	Unauthorized Privacy Violations	8 × 9 72	●	●	●	●	●	●	●	●	●	●	●	●	●
42	Types of Sensitive Data														
	L2: Privacy														
45	Other Unlawful/ Criminal Activities	4	●	●	●	●	●	●	●	●	●	●	●	●	●
	L2: Criminal Activities														

Table 8: The 7 most widely specified risk categories at level-3 across AI companies' policies.

Most Common Risk Categories. Table 8 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 .

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).

Least Common Risk Categories. Table 9 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.

	L3-Name	L4-Total													
16	Offensive Language	2	○	○	○	○	○	○	○	○	○	○	●	○	●
	L2: Hate/Toxicity														
19	Non-Consensual Nudity	1	○	○	○	○	○	○	○	○	○	●	○	○	○
	L2: Sexual Contents														
26	Deterring Democratic Participation	4	○	●	○	○	○	○	○	○	○	○	○	○	○
27	Disrupting Social Order	12	○	○	○	○	○	○	○	○	○	○	●	○	●
	L2: Political Usage														
29	Unfair Market Practices	2	○	○	○	○	○	○	○	○	○	○	●	○	●
30	Disempowering Workers	4	○	○	○	○	○	○	○	○	○	○	○	○	○
31	Fraudulent Schemes	2	●	○	●	○	○	○	○	○	○	○	○	○	○
	L2: Economic Harm														

Table 9: The 7 least often mentioned risk categories at level-3 across corporate AI policies.

C 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 [51]. 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.

C.1 Breakdown of Risk Categories by Jurisdiction

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 [10] primarily employing level-3 risk categories, whereas China's regulations [21, 22, 23, 59, 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.

C.1.A European Union

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

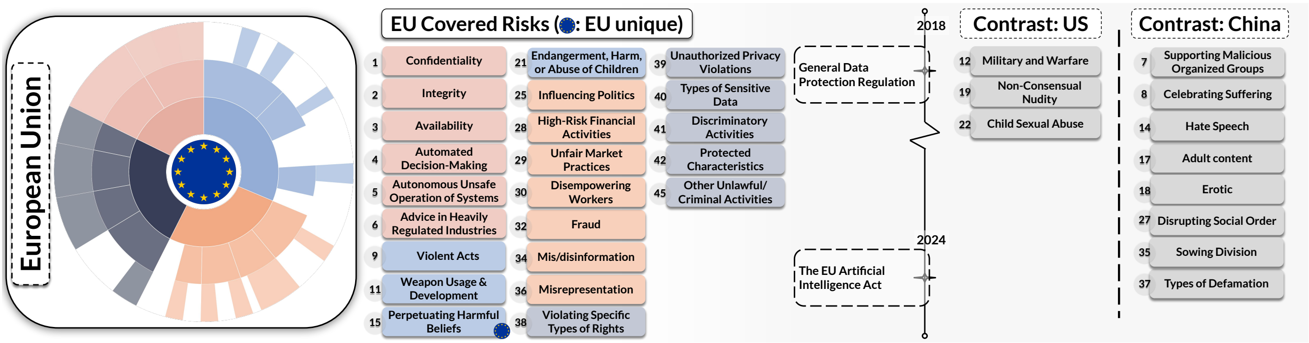

European Union
EU Covered Risks (●: EU unique)
1 Confidentiality
2 Integrity
3 Availability
4 Automated Decision-Making
5 Autonomous Unsafe Operation of Systems
6 Advice in Heavily Regulated Industries
9 Violent Acts
11 Weapon Usage & Development
15 Perpetuating Harmful Beliefs
21 Endangerment, Harm, or Abuse of Children
25 Influencing Politics
28 High-Risk Financial Activities
29 Unfair Market Practices
30 Disempowering Workers
32 Fraud
34 Mis/disinformation
36 Misrepresentation
38 Violating Specific Types of Rights
39 Unauthorized Privacy Violations
40 Types of Sensitive Data
41 Discriminatory Activities
42 Protected Characteristics
45 Other Unlawful/ Criminal Activities
2018
General Data Protection Regulation
2024
The EU Artificial Intelligence Act
Contrast: US
12 Military and Warfare
19 Non-Consensual Nudity
22 Child Sexual Abuse
Contrast: China
7 Supporting Malicious Organized Groups
8 Celebrating Suffering
14 Hate Speech
17 Adult content
18 Erotic
27 Disrupting Social Order
35 Sowing Division
37 Types of Defamation

Figure 3: The risks included in the GDPR and EU AI Act mapped as 23 level-3 categories in the AIR 2024. This includes risks described in recitals in the AI Act; see main text for discussion of limitations with respect to such risks.

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 [61, 13, 32, 11, 36, 41, 42]. 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 nonbinding provisions that explain the objectives of the law [47, 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 [13, 48].

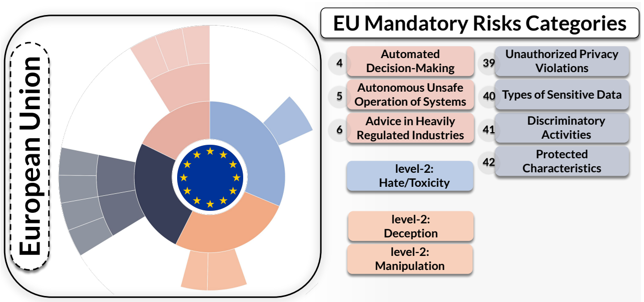

European Union
EU Mandatory Risks Categories
4 Automated Decision-Making
5 Autonomous Unsafe Operation of Systems
6 Advice in Heavily Regulated Industries
level-2: Hate/Toxicity
level-2: Deception
level-2: Manipulation
39 Unauthorized Privacy Violations
40 Types of Sensitive Data
41 Discriminatory Activities
42 Protected Characteristics

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.

C.1.B United States

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 [10]. The Executive Order is based in part on the voluntary National Institute of Standards and Technology AI Risk Management Framework [62] issued in January 2023, which has also inspired many state-level regulatory proposals [77]. 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 [54]. The aims of the Executive Order also include promoting innovation and competition, supporting

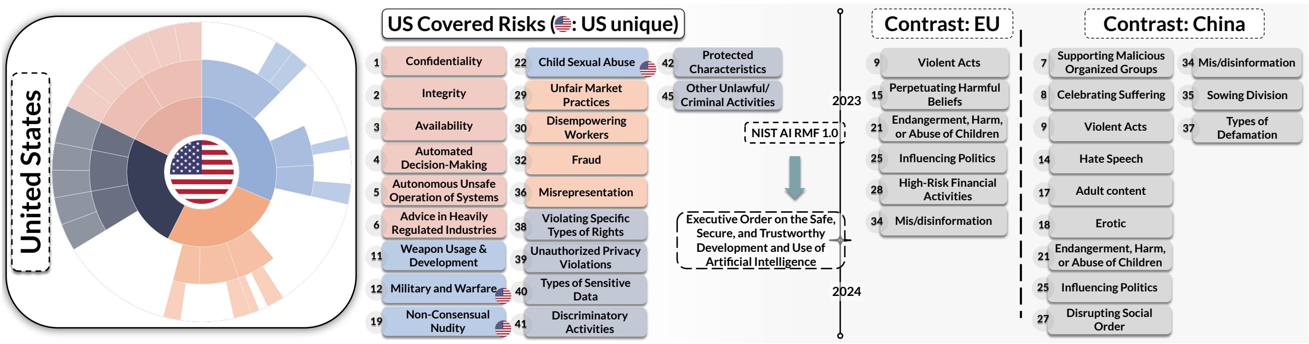

United States
US Covered Risks (🇺🇸: US unique)
1 Confidentiality
2 Integrity
3 Availability
4 Automated Decision-Making
5 Autonomous Unsafe Operation of Systems
6 Advice in Heavily Regulated Industries
11 Weapon Usage & Development
12 Military and Warfare
19 Non-Consensual Nudity
22 Child Sexual Abuse
29 Unfair Market Practices
30 Disempowering Workers
32 Fraud
36 Misrepresentation
38 Violating Specific Types of Rights
39 Unauthorized Privacy Violations
40 Types of Sensitive Data
41 Discriminatory Activities
42 Protected Characteristics
45 Other Unlawful/ Criminal Activities
2023
NIST AI RMF 1.0
Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
2024
Contrast: EU
9 Violent Acts
15 Perpetuating Harmful Beliefs
21 Endangerment, Harm, or Abuse of Children
25 Influencing Politics
28 High-Risk Financial Activities
34 Mis/disinformation
Contrast: China
7 Supporting Malicious Organized Groups
8 Celebrating Suffering
9 Violent Acts
14 Hate Speech
17 Adult content
18 Erotic
21 Endangerment, Harm, or Abuse of Children
25 Influencing Politics
27 Disrupting Social Order
34 Mis/disinformation
35 Sowing Division
37 Types of Defamation

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

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.

C.1.C China (mainland)

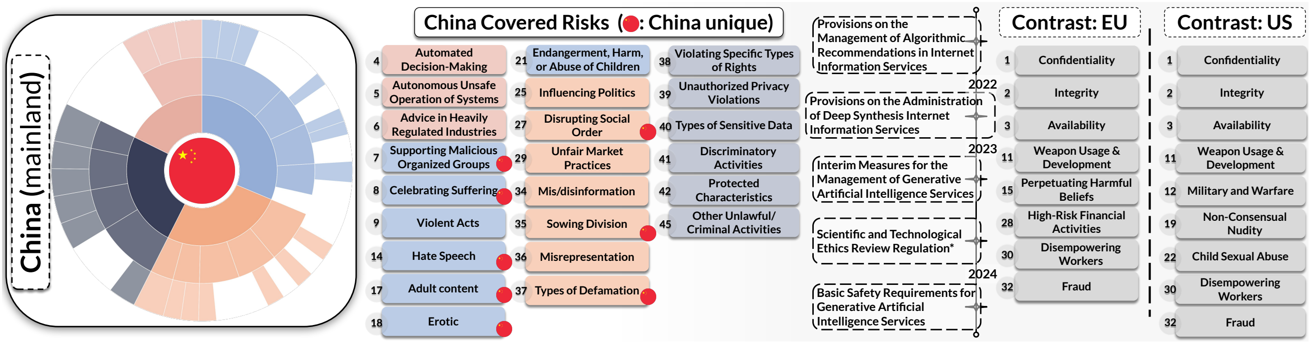

China (mainland)
China Covered Risks (●: China unique)
4 Automated Decision-Making
5 Autonomous Unsafe Operation of Systems
6 Advice in Heavily Regulated Industries
7 Supporting Malicious Organized Groups
8 Celebrating Suffering
9 Violent Acts
14 Hate Speech
17 Adult content
18 Erotic
21 Endangerment, Harm, or Abuse of Children
25 Influencing Politics
27 Disrupting Social Order
29 Unfair Market Practices
34 Mis/disinformation
35 Sowing Division
36 Misrepresentation
37 Types of Defamation
38 Violating Specific Types of Rights
39 Unauthorized Privacy Violations
40 Types of Sensitive Data
41 Discriminatory Activities
42 Protected Characteristics
45 Other Unlawful/ Criminal Activities
Provisions on the Management of Algorithmic Recommendations in Internet Information Services
2022
Provisions on the Administration of Deep Synthesis Internet Information Services
2023
Interim Measures for the Management of Generative Artificial Intelligence Services
Scientific and Technological Ethics Review Regulation*
2024
Basic Safety Requirements for Generative Artificial Intelligence Services
Contrast: EU
1 Confidentiality
2 Integrity
3 Availability
11 Weapon Usage & Development
15 Perpetuating Harmful Beliefs
28 High-Risk Financial Activities
30 Disempowering Workers
32 Fraud
Contrast: US
1 Confidentiality
2 Integrity
3 Availability
11 Weapon Usage & Development
12 Military and Warfare
19 Non-Consensual Nudity
22 Child Sexual Abuse
30 Disempowering Workers
32 Fraud

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 [86, 2, 87, 83, 43, 71, 80, 74, 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) [59], 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 [73].

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.

C.2 Comparative Analysis of Shared Generative AI Risk Categories in the Public Sector

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.

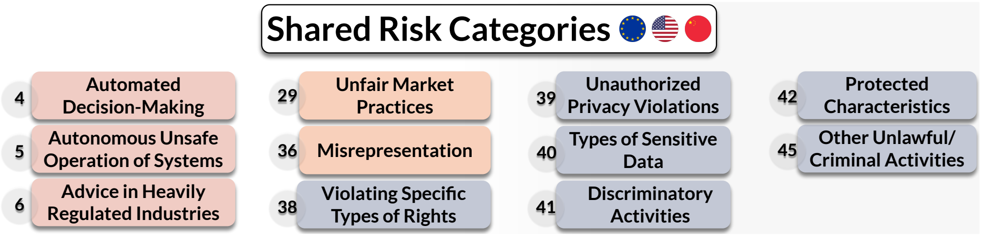

Shared Risk Categories
4
Automated Decision-Making
5
Autonomous Unsafe Operation of Systems
6
Advice in Heavily Regulated Industries
29
Unfair Market Practices
36
Misrepresentation
38
Violating Specific Types of Rights
39
Unauthorized Privacy Violations
40
Types of Sensitive Data
41
Discriminatory Activities
42
Protected Characteristics
45
Other Unlawful/ Criminal Activities

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

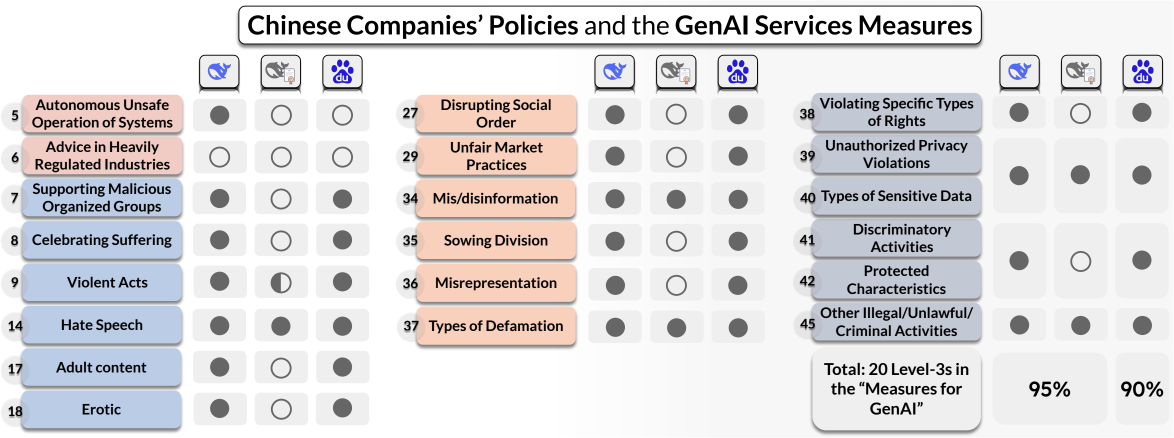
Interestingly, a closer examination of the level-4 risk categories within these shared level-3 categories reveals significant overlap in the specific risks considered by each jurisdiction. For example, within the Automated Decision-Making category, all three jurisdictions specify risks related to algorithmic 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 [50].

D Discussion

D.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.



Chinese Companies' Policies and the **GenAI Services Measures**

5	Autonomous Unsafe Operation of Systems	●	○	○
6	Advice in Heavily Regulated Industries	○	○	○
7	Supporting Malicious Organized Groups	●	○	●
8	Celebrating Suffering	●	○	●
9	Violent Acts	●	◐	●
14	Hate Speech	●	●	●
17	Adult content	●	○	●
18	Erotic	●	○	●
27	Disrupting Social Order	●	○	●
29	Unfair Market Practices	●	○	●
34	Mis/disinformation	●	●	●
35	Sowing Division	●	○	●
36	Misrepresentation	●	○	●
37	Types of Defamation	●	●	●
38	Violating Specific Types of Rights	●	○	●
39	Unauthorized Privacy Violations	●	●	●
40	Types of Sensitive Data			
41	Discriminatory Activities	●	○	●
42	Protected Characteristics			
45	Other Illegal/Unlawful/ Criminal Activities	●	●	●
	Total: 20 Level-3s in the "Measures for GenAI"	95%		90%

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 [8, 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."

D.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 assist policymakers or policy researchers in industry when drafting future AI policies. Without a greater level of detail in discussions of AI risk, it is difficult to understand the fact that superficial alignment between policies on broader, level-2 risk categories may not be reflective of alignment across policies with respect to 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 [12] 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 [14, 2]. 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.[3] We intend to analyze policies from a larger number of countries in future work.

E 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.

[3] 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] 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].

[3] Nouf Alfawzan, Markus Christen, Giovanni Spitale, and Nikola Biller-Andorno. Privacy, data sharing, and data security policies of women's mhealth apps: Scoping review and content analysis. *JMIR Mhealth Uhealth*, 10(5):e33735, 2022.

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

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

[10] 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.

[11] 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.

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

[13] 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.

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

[15] Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory. The long fuse: Misinformation and the 2020 election, 2021. Stanford Digital Repository: Election Integrity Partnership. v1.3.0.

[16] Inyoung Cheong, Aylin Caliskan, and Tadayoshi Kohno. Safeguarding human values: rethinking us law for generative ai's societal impacts. *AI and Ethics*, pages 1–27, 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-standards-in-ai/, 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] Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. *Berkman Klein Center Research Publication*, (2020-1), January 2020. Available at SSRN: https://ssrn.com/abstract=3518482 or http://dx.doi.org/10.2139/ssrn.3518482.

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

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

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

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

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

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

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

[49] 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.

[50] 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].

[51] 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.

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

[53] 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.

[54] 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.

[55] Meta. Meta llama-2 and llama-3's acceptable use policy. https://ai.meta.com/llama/use-policy/, 2023.

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

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

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

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

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

[61] 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].

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

[63] Future of Life Institute. Can we rely on information sharing? examining gpais developers' terms of use, October 2023.

[64] 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.

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

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

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

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

[69] 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.

[70] 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.

[71] 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.

[72] Megan Shahi, Adam Conner, Nicole Alvarez, and Sydney Bryant. Generative ai should be developed and deployed responsibly at every level for everyone. Center for American Progress, 2024.

[73] 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].

[74] 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].

[75] 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.

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

[77] 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.

[78] The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-int
2023.

[79] 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.

[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. https://digichina.stanford.edu/work/how-will-chinas-generative-ai-regulations-shape-the-future-a-digichina-forum/, 2023. [Online; accessed 2-Jun-2024].

[81] Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*, 2024.

[82] 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.

[83] 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].

[84] 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.

[85] 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.

[86] 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.

[87] 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].

[88] 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.