Generative AI Risk Categorization Decoded: Comparing Public and Private Sector Policies

Anonymous Authors [1]

As generative AI systems have been deployed at an increasingly frenetic pace, researchers have prioritized identifying the risks and harms of such systems (Weidinger et al., 2021; Shelby et al., 2023; Hoffmann & Frase, 2023). 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 derived from 16 policies from foundation model developers as well as eight government policies from the European Union (European Commission, 2024; European Parliament & Council of the European Union, 2016), United States (The White House, 2023), and the People's Republic of China (Cyberspace Administration of China, 2023; 2024; 2022; Ministry of Science and Technology of China, 2023; Cyberspace Administration of China, 2021). 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.

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 (Bommasani et al., 2024). *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 (Anthropic, 2023), DeepSeek (DeepSeek, 2024)), others use broad language that explicitly addresses just a handful of risks (e.g. Mistral (Mistral, 2024)), illustrating different legal strategies adopted across the private sector (of Life Institute, 2023). *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 (Zhang, 2024; Toner et al., 2023). *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 public and private sector policies. We plan to release all data associated with our taxonomy to facilitate other efforts to build on this work after the anonymous review period.

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 (Vidgen et al., 2024; Zou et al., 2023; Qi et al., 2024; Li et al., 2024). For example, Vidgen et al. (2024) 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 these 7 categories. Our taxonomy includes 45 "level-3" risk categories at a comparable level of specificity (see Figure 1). In addition, existing taxonomies are somewhat disconnected from what many companies do in practice. Klyman (2024) 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 included in 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 (Shahi et al., 2024). By including the risks from such policies (as well as firms' platform-wide policies) in tandem with 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

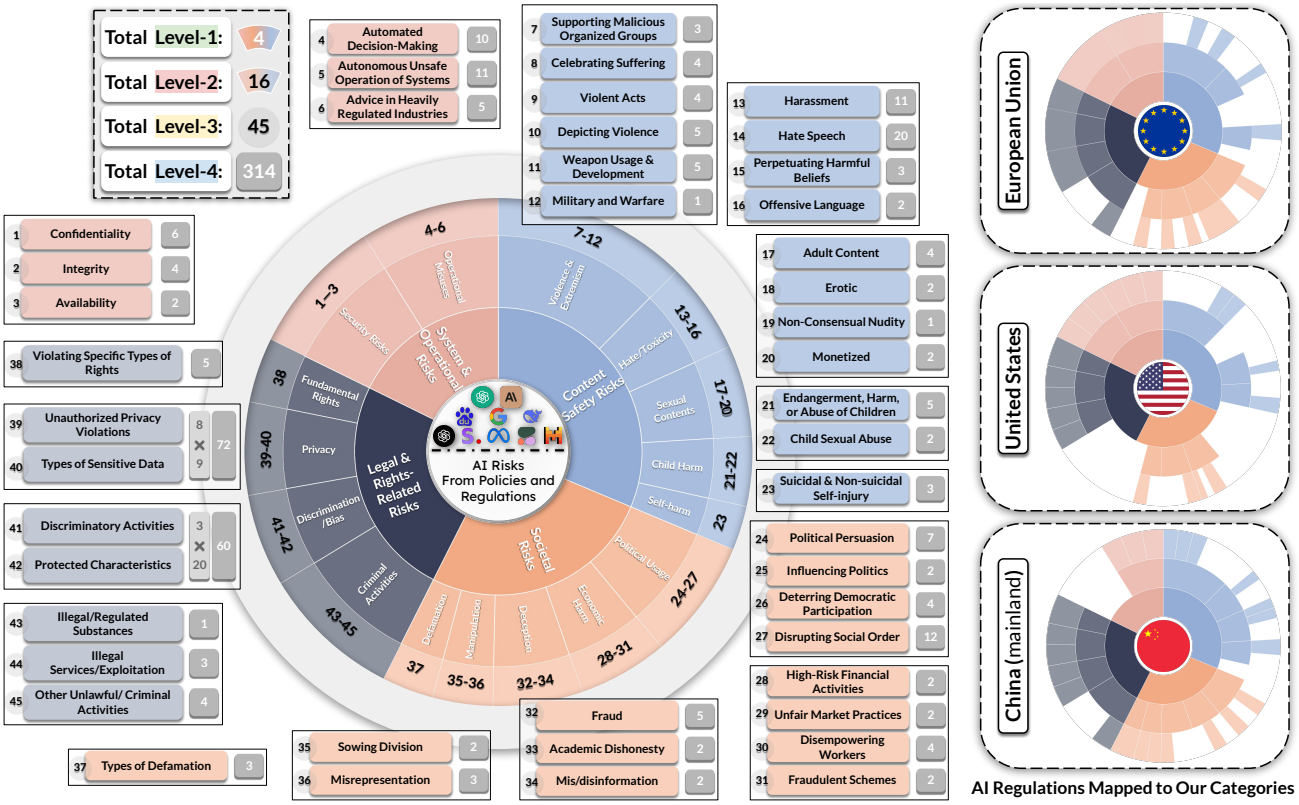

Figure 1: Overview of the generative AI risk taxonomy, derived from 24 policy and regulatory documents, which includes 314 unique risk categories. Charts on the right-hand side correspond to major AI regulations in the EU, US, and China.

categories in policies from governments and companies (Mayring, 2015). This was done inductively (Elo & Kyngäs, 2008), with categories drawn directly from such policies without major alterations. Following on prior work in this area (Klyman, 2024), we drew on similar methods used to categorize AI ethics guidelines (Fjeld et al., 2020), privacy policies (Alfawzan et al., 2022), and content moderation guidelines (Center for an Informed Public et al., 2021).

Comparing companies' policies to those of governments lends significant insight. At a high level, DeepSeek's platform-wide policies (DeepSeek, 2024; 2023) 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 than 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 (The White House, 2023) specifies "Non-consensual Nudity" as a risk, but only one company policy mentions it (Stability, 2024). Legal and rights-related risks, such as discrimination, privacy violations, and illegal activities, are the most consistent risk categories across both government regulations and company policies. We find greatest variance across jurisdictions on Content Safety Risks, reflecting different cultural values (e.g. related to sexual content). Political risks also vary significantly, with firms tailoring their policies to the regulations of the regions where they operate.

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 (Cyberspace Administration of China, 2023), 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 (Baidu, 2023) 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 (Sheehan, 2024).

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.

References

Alfawzan, N., Christen, M., Spitale, G., and Biller-Andorno, N. Privacy, data sharing, and data security policies of women's mhealth apps: Scoping review and content analysis. *JMIR Mhealth Uhealth*, 10(5):e33735, 2022. doi: 10.2196/33735. URL https://mhealth.jmir.org/2022/5/e33735.

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

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

Bommasani, R., Klyman, K., Longpre, S., Xiong, B., Kapoor, S., Maslej, N., Narayanan, A., and Liang, P. Foundation model transparency reports, 2024.

Center for an Informed Public, Digital Forensic Research Lab, Graphika, and Stanford Internet Observatory. The long fuse: Misinformation and the 2020 election, 2021. URL https://purl.stanford.edu/tr171zs0069. Stanford Digital Repository: Election Integrity Partnership. v1.3.0.

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

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

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

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

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

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

Elo, S. and Kyngäs, H. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008. doi: https://doi.org/10.1111/j.1365-2648.2007.04569.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2648.2007.04569.x.

European Commission. The eu artificial intelligence act, 2024. URL https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.

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.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., and Srikumar, M. 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.

Hoffmann, M. and Frase, H. Adding structure to ai harm: An introduction to cset's ai harm framework. Technical report, Center for Security and Emerging Technology, July 2023. URL https://doi.org/10.51593/20230022.

Klyman, K. Acceptable use policies for foundation models: Considerations for policymakers and developers. Stanford Center for Research on Foundation Models, April 2024. URL https://crfm.stanford.edu/2024/04/08/aups.html.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Mayring, P. *Qualitative Content Analysis: Theoretical Background and Procedures*, pp. 365–380. Springer Netherlands, Dordrecht, 2015. ISBN 978-94-017-9181-6. doi: 10.1007/978-94-017-9181-6_13. URL https://doi.org/10.1007/978-94-017-9181-6_13.

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

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

of Life Institute, F. Can we rely on information sharing? examining gpais developers' terms of use, October 2023. URL https://futureoflife.org/ai-policy/can-we-rely-on-information-sharing/.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Shahi, M., Conner, A., Alvarez, N., and Bryant, S. Generative ai should be developed and deployed responsibly at every level for everyone. Center for American Progress, 2024. URL https://www.americanprogress.org/wp-content/uploads/sites/2/2024/02/GenerativeAI-report.pdf.

Sheehan, M. 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].

Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla, N., Gallegos, J., Smart, A., Garcia, E., and Virk, G. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction, 2023.

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

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-intel 2023.

Toner, H., Haluza, Z., Luo, Y., Dan, X., Sheehan, M., Huang, S., Chen, K., Creemers, R., Triolo, P., and Meinhardt, C. How will china's generative ai regulations shape the future? a digichina forum, April 19 2023.

Vidgen, B., Agrawal, A., Ahmed, A. M., Akinwande, V., Al-Nuaimi, N., Alfaraj, N., Alhajjar, E., Aroyo, L., Bavalatti, T., Blili-Hamelin, B., et al. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*, 2024.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Zhang, A. H. 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.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.