Risk-Based and Proportionate Regulation of Generative AI: A Study on Content Moderation, Disinformation, and Cybersecurity

Yuanquan Meng*

Capital University of Economics and Business, Beijing, China

*mengyuanquan27@163.com

Abstract. The governance challenges of generative AI are increasing in content moderation, disinformation, and cybersecurity fields. This study develops a dual-dimensional analytical framework combining risk tiering and proportionality assessment to assess the effectiveness of regulations from four major jurisdictions (EU, US, China, the UK). Through comparative analysis using the proportionality assessment tool (suitability, necessity, and proportionality stricto sensu), the study reveals a descending order of regulatory effectiveness from the cybersecurity domain to disinformation. This study reveals that proportionality effectiveness is limited by the quantifiability of the risk and the complexity of value tension. The cybersecurity domain shows the highest proportionality effectiveness because of the well-defined technical indicators and fewer conflicts of rights. In contrast, the disinformation domain is challenged by the unclear risk boundaries and the complexity of freedom of expression. Current frameworks exhibit notable progress in risk classification, particularly the EU AI Act's four-tier system, yet structural deficiencies persist in systematic proportionality application. This study provides a novel proportionality assessment tool for evaluating the effectiveness of the regulations and proposes domain-specific strategies to improve proportionality effectiveness.

Keywords: Generative AI, Risk-based regulation, Proportionality principle, Content moderation, Disinformation, Cybersecurity, AI governance.

1 Introduction

When deepfake videos are disseminated on a massive scale during election cycles, AI-generated phishing emails bypass conventional security defenses, and automated content factories produce harmful information at minimal cost, a fundamental question emerge: whether existing regulatory systems possess the capacity to address the systemic risks posed by generative AI. Generative AI systems represented by large language models and diffusion models have given rise to urgent and interconnected challenges in three key domains, namely content moderation, disinformation, and cybersecurity [1]. These challenges have three significant structural characteristics that make traditional regulatory systems increasingly ineffective [2]. These character -

D. Magni et al. (eds.), *Proceedings of the 2026 3rd International Conference on Applied Economics, Management Science and Social Development (AEMSS 2026)*, Advances in Economics, Business and Management Research 389,
https://doi.org/10.2991/978-94-6239-672-2_46

istics are scalability, international reach, and the dynamic nature of technology. In response to these challenges, major jurisdictions around the globe are pursuing different regulatory strategies. Nevertheless, one significant factor that dominates all of them is how to control risks while avoiding undue constraints on technology innovation and fundamental human rights [3]. Research has revealed a "race to regulation" between nations in the context of AI governance; the effectiveness of risk control and the need to maintain innovation space are in tension ever since. Moreover, blanket regulatory approaches are increasingly criticized for failing to account for the diverse risk profiles across different application contexts [4].

This predicament has led to the emergence of two different logics of regulatory design that are of equal interest to both academic theory and policy. The risk-based logic is oriented towards the need to identify and prioritize risks through the classification of AI systems according to the severity of the harm that might occur. On the other hand, the logic of proportionality is oriented towards the need to address the question of regulatory adequacy and intrusiveness by requiring that the proportionality of the regulatory response to the risks [5]. An analysis of the EU AI Act reveals that the four-tier risk classification system is a legislative codification of the logic of proportionality that aims to attain the optimal balance between the need to promote innovation and the need to address the risks [6]. From the broader theoretical perspective of the regulation of AI ethics, the gap between principles and practices offers an important reference point for the assessment of the actual effectiveness of current regulatory frameworks. However, the current gap in the literature is that the two logics of regulatory design are not integrated into a unified analytical framework for the regulation of generative AI risks [7]. The global digital governance landscape has been characterized as a competitive arena among multiple "digital empires", yet how jurisdictional regulatory divergences manifest at the level of proportionality application remains underexamined.

To address these gaps, this study develops a dual-dimensional analytical framework that integrate risk tiering and proportionality assessment, and conducts a systematic comparative analysis of existing regulatory measures across the three risk domains to reveal their underlying risk classification logics, regulatory design features, and proportionality performance. The research is underpinned by two research questions. RQ1 aims to analyze how major regulatory frameworks classify generative AI-related risks and design regulatory responses to such risks in the domains of content moderation, disinformation, and cybersecurity. RQ2 aims to examine the regulatory measures in the domains through the proportionality test of suitability, necessity, and proportionality stricto sensu.

2 Research Design

2.1 Analytical Framework

On the basis of these two core challenges in the design of a regulatory regime, as identified in the introduction, this study develops a matrix-based analytical framework

that integrates a Risk Tiering dimension with a 'Proportionality Assessment' dimension.

The risk tiering dimension draws on the four-tier risk classification logic of the EU AI Act, which groups AI systems into four tiers based on the potential severity of harm, namely unacceptable risk, high risk, limited risk, and minimal risk. This study applies this classification logic to the three risk domains of generative AI, establishing risk-level mapping criteria for content moderation, disinformation, and cybersecurity, with particular attention to whether different regulatory frameworks assign consistent risk tiers to the same domains and the underlying logics behind any divergences.

The proportionality assessment dimension adopts the tripartite test rooted in the European public law tradition. The suitability test evaluates whether a regulatory measure is capable of achieving its stated regulatory objective. The necessity test examines whether less restrictive alternative measures exist. The proportionality stricto sensu test assesses whether the benefits of a regulatory measure are commensurate with the burdens it imposes. By subjecting the regulatory measures of the three risk domains to the tripartite test, the framework can identify the proportionality characteristics of each domain. The framework structure is shown in Fig. 1.

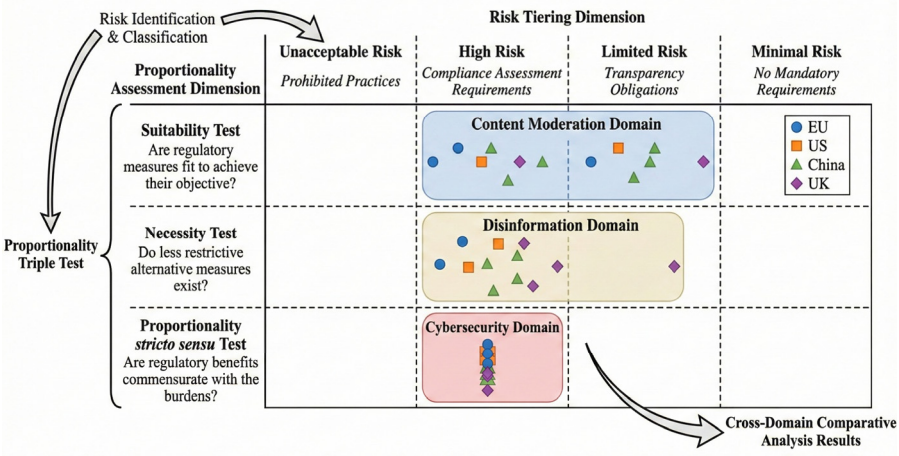

Fig. 1. Dual-Dimensional Analytical Framework: Risk Tiering × Proportionality Assessment.

2.2 Data Collection and Selection Criteria

The policy text selection covers core regulatory instruments from four major jurisdictions, including the EU AI Act (2024), US Executive Order on Safe, Secure, and Trustworthy AI (2023), China's Interim Measures for the Management of Generative AI Services (2023), and the UK's Pro-Innovation Approach to AI Regulation White Paper (2023), along with the OECD AI Principles as an international reference. The rationale for the selection of these policies is based on the fact that they represent the three major types of regulatory approaches—legislative, executive order-based, and

principle-based — thereby making them representative for cross-jurisdictional comparative analysis.

Academic literature. The academic literature selection includes sources from Scopus and Google Scholar, covering international peer-reviewed journals published between 2020 and 2025. The literature selection criteria include sources containing explicit analysis of regulatory mechanisms or discussions of policy design, excluding purely technical papers, conference papers, and opinion-based commentaries.

2.3 Analytical Procedure

The analytical process comprises three stages. Thematic coding employs a deductive-inductive hybrid approach in which initial codes are derived from the theoretical framework (risk categories, regulatory instruments, proportionality criteria), while codes are also developed to capture jurisdiction-specific nuances that are not initially considered in the theoretical framework. Each policy document is coded to identify relevant criteria for determining risks, different types of regulatory instruments, and mechanisms for enforcing each type of policy, corresponding to each of the three risk domains. To further establish coding reliability, a portion (20%) of the coding is cross-checked by a second researcher with knowledge of relevant literature on AI governance, resulting in an excellent agreement rate of 87% (Cohen's kappa = 0.82).

The framework mapping systematically placed the coding results on the dual-dimensional analytical structure that was developed above. This created a structured analytical matrix that reflected the intersection of the three risk domains and the two analytical dimensions. Divergences in the classification of risk or the regulatory approach were identified.

The proportionality testing involved applying the tripartite proportionality test to the principal regulatory measures in individual risk domains. To operationalize the proportionality testing, domain-specific evaluation criteria were developed through which the suitability of the measures was evaluated on the basis of the empirical evidence on the alignment of the measures with the objective of the regulation, the necessity of the measures was evaluated on the basis of the evidence on the existence of alternative measures that were less restrictive, and proportionality stricto sensu was evaluated on the basis of the stakeholder impact assessments and the regulatory cost-benefit analyses. Each of the measures was rated on a three-point scale based on the strength of evidence for their proportionality performance. Although the process is necessarily evaluative and judgment-based, the structured process and the dual coding process enhance its transparency and reproducibility of the process. The results of the three analytical stages are presented in the following section.

3 Results

3.1 Risk Profiling and Regulatory Responses Across Three Domains

The challenges that generative AI poses to content moderation systems stem primarily from the dramatic expansion in the scale of harmful content production and the adver-

sarial evasion of existing moderation mechanisms by generative AI technologies. In response to these risks, different jurisdictions have developed differentiated regulatory approaches. The EU AI Act classifies relevant AI systems from high risk to limited risk depending on the application context, imposing transparency obligations and AI-generated content labeling requirements. China requires service providers to complete security assessments and file records prior to launch. The US relies predominantly on voluntary industry commitments without establishing a mandatory risk classification mechanism. The UK draws on its Online Safety Act framework and grants sectoral regulators domain-specific adaptable regulatory authority. Despite these divergent classification pathways, human review and user complaint mechanisms appear as common instruments across multiple frameworks, reflecting a basic consensus among jurisdictions on the logic of human-machine collaborative moderation at the enforcement level.

If the jurisdictional difference in content moderation policies is best evidenced by the level of detail used to classify risks, then disinformation policies reveal an even more fundamental level of disagreement between jurisdictions. The creation of deepfake media and the automated dissemination of false narratives represent the core risks in this domain, yet there is a fundamental difference in how each jurisdiction defines the scope or trigger of this risk. The EU's combined framework of the AI Act and DSA treats disinformation as part of the high-risk category. The US's current policy includes an ambiguous definition of disinformation, which is heavily influenced by the need to protect freedom of speech as defined by the First Amendment. The current policy in China includes a mechanism combining content tracking and platform liability, while no such rules have yet been implemented in the UK. However, such a difference is not merely a product of technical differences but is instead based upon each jurisdiction's normative hierarchy of freedom of expression, information security, and governance. Within this context, AI-generated content provenance labeling, platform liability, and election protections represent the core policy mechanisms currently available to each jurisdiction.

In stark contrast to the divergent landscape of the previous two domains, the cybersecurity domain shows the highest degree of cross-jurisdictional consensus among the three risk domains. The increasing sophistication of AI-assisted cyber-attack capabilities, the development of automated vulnerability detection and exploitation techniques, and the evolution of sophisticated social engineering attacks have been rated as high-risk in all prominent risk frameworks. The EU AI Act has established stringent security assessment requirements, the US Executive Order has mandated the use of red-teaming and safety assessments for frontier models, China has mandated the use of cybersecurity classification and protection standards, and the UK has had a security guidance framework for AI systems issued by the National Cyber Security Centre. The development of this consensus is strongly correlated with the quantifiability of cybersecurity risks, which can be objectively assessed by metrics such as the number of detected vulnerabilities and the success rate of attacks. A comparative overview of the risk profiles and the associated regulations in the three domains is provided in Table 1.

Table 1. Risk Profiling Matrix Across Three Domains and Four Jurisdictional Frameworks.

Risk Domain	Key Risk Manifestations	EU AI Act	US EO 14110	China Interim Measures	UK AI White Paper
Content Moderation	Harmful content generation; adversarial evasion of safety filters	High-risk to limited-risk (tiered)	Voluntary commitments; industry self-regulation	Pre-release security assessment; content labeling obligations	Principle-based; cross-sectoral guidance
Disinformation	Deepfake generation; automated false narrative production	High-risk (AI Act + DSA synergy)	Ambiguous classification; First Amendment constraints	Provenance tracking; platform liability for dissemination	No binding classification; voluntary code of practice
Cybersecurity	AI-assisted attack automation; vulnerability exploitation at scale	High-risk (explicit classification)	High-risk; mandatory red-teaming for dual-use models	High-risk; security review and incident reporting	High-risk; alignment with NCSC guidelines

3.2 Proportionality Assessment Results

The test of suitability suggests that all of the regulatory instruments in each of the domains share an intrinsic, albeit varying, level of logical connection to their respective objectives, though this level of connection decreases in a gradient from cybersecurity to content moderation and then to disinformation. In the cybersecurity domain, the most direct level of connection between risk mitigation objectives and technically oriented standards-based measures is evident. In the content moderation domain, mechanisms for pre-filtering content face significant technical hurdles related to the proliferation of adversarially generated content. In the disinformation domain, mechanisms for provenance labeling confront fundamental questions of enforceability in an environment where open-source models of this type of content are ubiquitous.

The test of necessity continues and expands upon this trend of differentiation between the domains. In the cybersecurity domain, mechanisms for red teaming and security assessments share a high level of functional necessity due to their intrinsic technical difficulty of replacement by less restrictive measures. In the content moderation domain, the presence of mechanisms for mandatory pre-filtering and the combined mechanism of post-hoc labeling and user reporting create an environment in which an evaluation of necessity is possible, though one in which there is a significant level of variation between the restrictiveness of the mechanisms and measures at play. In the disinformation domain, the question of necessity may be seen as one of fundamental difficulty, given the ambiguous definition of the term "disinformation".

The proportionality stricto sensu test takes the analysis to the next level, i.e., the level of value trade-offs, thereby further emphasizing the fundamental differences between the three domains. The balance between the costs and benefits of regulations

is most transparent in the cybersecurity domain, as costs and benefits can be measured. The content moderation domain ranks second in terms of transparency, as it includes both technical measurability and value trade-offs regarding freedom of speech. The disinformation domain includes significant value trade-offs between freedom of expression and information security, making the proportionality stricto sensu assessment complexity the highest among the three domains. Based on the results of the tripartite test, a summary of the assessment findings is provided in Table 2. The overall proportionality performance is best in the cybersecurity domain, worst in the disinformation domain, and intermediate in the content moderation domain.

Table 2. Proportionality Assessment Matrix: Tripartite Test Results by Risk Domain.

Proportionality Test	Content Moderation	Disinformation	Cybersecurity
Suitability	Moderate: technical filters partially effective but prone to over-blocking and context insensitivity	Low: enforceability questioned due to definitional ambiguity and rapid content mutation	High: direct alignment between technical standards and risk mitigation objectives
Necessity	Moderate: clear alternative measures exist (user reporting, contextual moderation, graduated enforcement)	Low: absence of baseline definitions undermines necessity assessment; no agreed threshold for intervention	High: limited less-restrictive alternatives; technical mandates represent minimum viable intervention
Stricto Sensu Proportionality	Moderate: tension between speech rights and safety yields contested cost-benefit balance	Low: deepest value trade-offs between expression freedom and information integrity; highest regulatory uncertainty	High: clearest cost-benefit balance; limited fundamental rights conflicts; broad stakeholder consensus
Overall Assessment	Intermediate	Weakest	Strongest

3.3 Cross-Domain Regulatory Coherence

From the domain-specific analyses provided above, some interesting cross-domain patterns emerge. For one, transparency requirements, as the only cross-domain instrument shared among the three domains, still lack cross-domain coordination in their implementation, suggesting that these frameworks continue to operate in a domain-siloed manner to address what is, in essence, cross-domain risk. In terms of consistency in the classification of risks, a clear gradient is observable, with cybersecurity faring best, followed by content moderation, which still heavily depends on determinations made in individual cases, while disinformation performs worst in this regard. Yet another interesting pattern is that, despite some convergence in regulatory intensity, these jurisdictions still exhibit substantial divergences in addressing the same risk domain. The reason for this is not differences in technical capacity but the

continued influence of different legal traditions and their respective value orderings on regulatory design decisions.

4 Discussion

The analytical results obtained above suggest that the differences in the proportionality performance achieved across the three risk domains are not accidental but rather follow a systematic pattern related to the two factors identified. Thus, the cybersecurity domain, which is marked by clear indicators of risks and limited value tensions, is seen as the best performer among the three in the application of the tripartite proportionality test [8], due to the technically defined standards-based regulatory approach adopted in the field, which facilitates the evaluation of the suitability, necessity, and proportionality stricto sensu of regulatory interventions based on objective benchmarks [9]. At the opposite end of the spectrum is the disinformation domain, which is seen as the most challenging for proportionality evaluation due to the unclear boundaries of the risks involved and the close relationship with the protection of fundamental rights such as freedom of expression, while the content moderation domain occupies the middle ground. Again, these differences are not accidental but rather reflect the intrinsic difficulties faced by the regulatory design in addressing risks of different kinds, where the decrease in the quantifiability of the risks and the increase in the complexity of the value tensions involved lead to a corresponding decrease in proportionality performance. The general structure of the explanatory model based on the two factors identified is presented in Fig. 2. It is worth noting that the domain-siloed approach to security, ethics, and privacy issues in the literature exacerbates the challenges for regulatory coordination across the domains.

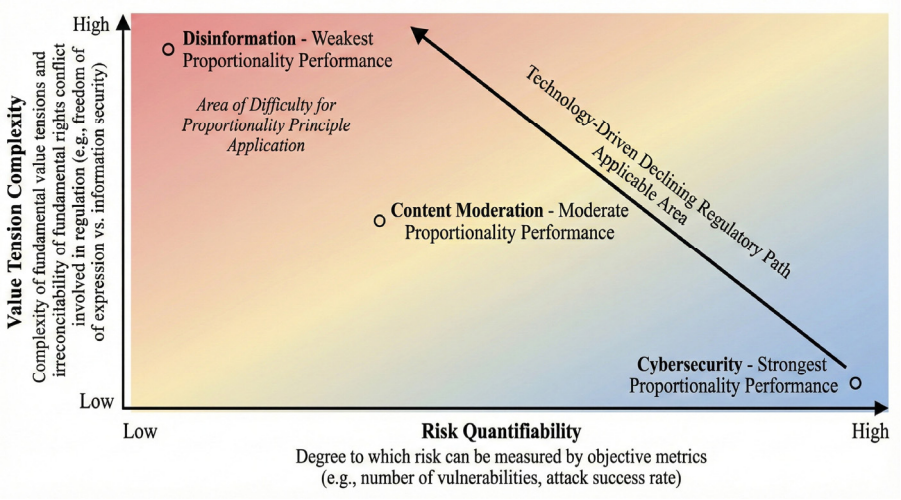

Fig. 2. Domain Positioning on Risk Quantifiability × Value Tension Complexity.

If the results of the current study are placed within the context of the broader academic discourse, they reveal meaningful points of convergence and extension in relation to the discussion on the global diffusion effects of the EU regulatory model. More specifically, in relation to the risk classification framework of the EU's AI Act, it has been observed in the literature that the framework in question has the potential to affect the regulatory framework of other jurisdictions, as evidenced by institutional diffusion dynamics [10]. The results of the comparative analysis of the current study corroborate the aforementioned assertion; however, they extend it by indicating that the degree of regulatory convergence varies across the different risk domains, ranging from the highest in the cybersecurity domain to the lowest in the disinformation domain, thus underscoring that the efficacy of institutional diffusion is shaped by the inherent characteristics of the risk domain in question. Finally, in relation to the effectiveness of the regulatory model in the context of the EU, scholars have argued that experimentalist governance offers a better explanation for the evolution of the global regulatory framework in the field of AI rather than the process of diffusion, as indicated in the literature, and the results of this study offer partial support for this literature. Finally, in relation to the effectiveness of the regulatory model in the context of the different risk domains, it should be noted that the proportionality tensions in the context of the content moderation domain, as well as the relationship between the latter and the disruptive effects of synthetic content on the public sphere, have been discussed in the literature, whereas the regulatory fragmentation in the disinformation domain has been further confirmed by the results of the empirical analysis of the legislation in the field of deepfake regulation, as well as tensions with the protection of human rights.

The above results have dual significance. On the theoretical level, the analysis confirms the validity of the dual-dimensional analytical method, as the sole use of risk tiering cannot be considered appropriate for assessing the adequacy of regulations, and the introduction of the proportionality principle as an analytical instrument is indispensable to avoid both regulatory excess and regulatory deficit. On the practical level, the results of the analytical process suggest that it is necessary to apply differentiated proportionality calibration approaches, depending on the specific characteristics of each risk domain, and to prioritize clarifying risk determination criteria rather than simply increasing the number of regulatory instruments in the domain of disinformation. This study has several limitations: it relies on policy texts rather than empirical evidence of the results of implementing regulations; it is limited to four major jurisdictions and does not consider regulations in developing countries; and it relies on proportionality assessments that require a degree of researcher interpretation.

5 Conclusion

Using a dual-dimensional analytical model based on the risk tiering and proportionality assessment, this study systematically conducts a comparative analysis of generative AI regulations across the three risk domains of content moderation, disinformation, and cybersecurity. The results of the analysis reveal that the proportionality perfor-

mance of the measures in the three risk domains is subject to the quantifiability of the risk and the complexity of the value conflicts, showing a descending gradient in proportionality performance from cybersecurity to content moderation and finally to disinformation. The cybersecurity risk domain achieves the optimal proportionality performance across the tripartite test due to the clear delineation of the technical indicators of risk and the limited value conflicts involved in the assessment of the AI measures in the domain. The content moderation risk domain achieves an intermediate level of proportionality performance due to the coexistence of the technical assessability of the risk and the value conflicts related to the right of free speech. The disinformation risk domain faces the most challenging proportionality test due to the unclear delineation of the risk boundaries and the close association with the fundamental right of freedom of expression. Although the current regulatory approaches have achieved considerable success in the risk classification of the AI measures, as reflected in the four-tier risk classification of the AI measures in the EU AI Act, which has become an important reference point for global AI governance–structural deficiencies in the application of the proportionality principle to the systematic assessment of AI regulation remain a key concern. This is reflected in the ambiguous risk determination criteria in the disinformation domain and the failure to adequately apply the necessity test (i.e., assessment in the absence of less restrictive alternatives) in the content moderation domain.

These contributions can be seen as operating at two levels. At the methodological level, the dual-dimensional approach integrates the tools of risk tiering and proportionality assessments within a reusable analytical instrument for assessing the appropriateness of AI regulatory interventions, addressing the shortcoming that a single risk classification perspective cannot adequately capture both regulatory excess and deficit. At the level of substantive findings, the cross-domain comparative analysis provides insights into the different demands that various types of risks place on the regulatory logic and design: the cybersecurity domain is well-suited for a technically informed standards-based regulatory design logic, the disinformation domain requires prioritizing the clarification of the definition of risks, and the content moderation domain requires a more refined balance of ex ante intervention and ex post remediation.

Future research can be extended in the following directions. The addition of stakeholder interviews or survey-based research could be considered to empirically validate the analytical framework developed through this research. This would compensate for the limitations of policy text analysis in assessing the implementation effectiveness. Additionally, as the EU AI Act takes full force in 2026, the proportionality performance of regulatory practices in different risk domains will provide valuable empirical evidence for dynamically calibrating the framework. Further research could be conducted to apply the analytical framework to other emerging risk domains, such as the labor market impact of generative AI or intellectual property. Finally, the addition of more regulatory practices from developing countries could be considered to improve the global representativeness of the analysis. This could provide a better comparative basis to develop a more inclusive global AI governance framework.

References

1. P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in Proc. 2023 ACM Conf. Fairness, Accountability, and Transparency (FAccT), Chicago, IL, USA, 2023, pp. 1112–1123.
2. A. Taeihagh et al., "Governance of generative AI," Policy and Society, vol. 44, no. 1, pp. 1–22, 2025.
3. N. A. Smuha, "From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence," Law, Innovation and Technology, vol. 13, no. 1, pp. 57–84, 2021.
4. M. C. Buiten, "Towards intelligent regulation of artificial intelligence," European Journal of Risk Regulation, vol. 10, no. 1, pp. 41–59, 2019.
5. M. Ebers, "Truly risk-based regulation of artificial intelligence: How to implement the EU's AI Act," European Journal of Risk Regulation, vol. 15, no. 4, pp. 652–677, 2024.
6. L. Floridi, The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities. Oxford, UK: Oxford University Press, 2023.
7. A. Bradford, Digital Empires: The Global Battle to Regulate Technology. Oxford, UK: Oxford University Press, 2023.
8. G. Fiorinelli and A. Ferretti, "Regulating AI to combat tech-crimes: Fighting the misuse of generative AI for cyber-attacks and digital offenses," Technology and Regulation, vol. 2025, pp. 1–18, 2025.
9. N. Sava et al., "Generative AI cybersecurity and resilience," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1568360, 2025.
10. E. Hine and L. Floridi, "The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market," arXiv preprint arXiv:2208.12645, 2022.