Advances in Humanities Research Vol.12 Issue 2 EWA Publishing
Available Online: 11 June 2025 DOI: 10.54254/2753-7080/2025.23849

The digital boundaries of free speech: legal interventions on hate speech and disinformation in the age of social media

Xun Zhang[1], Xiaofeng Cheng[2]*

[1]The University of Leeds, Leeds, England
[2]University of New South Wales, Sydney, Australia

*Corresponding Author. Email: rara481846778@gmail.com

Abstract. In the digital media era, extreme remarks and fake news on social platforms are constantly impacting the limits of freedom of expression. This study selects three jurisdictions—the European Union, the United States, and China—to compare and analyze the institutional development of online speech governance. By analyzing the practical conflict between the European Digital Services Act and the First Amendment to the U.S. Constitution, the paper reveals the value gap between protecting freedom of expression and implementing content control in different jurisdictions. Platform content audit data and post-removal appeal cases show that the existing governance system has structural problems such as unclear implementation standards and unbalanced allocation of audit resources. Especially in the interaction between algorithmic recommendation mechanisms and manual audits, users often encounter difficulties such as blocked appeal channels and opaque removal procedures. The research proposes the establishment of a hierarchical and classified content governance framework, the promotion of transnational platforms to establish a traceable audit log system, and the exploration of speech risk assessment models based on cultural context, so as to provide an institutional guarantee for the construction of a digital discourse space with equal rights and responsibilities.

Keywords: free speech, disinformation, hate speech, digital governance, social media regulation

1. Introduction

The social media process that is reconstructing the ecology of free expression in the 21st century is essentially the embodiment of the interplay between technological power and legal regulation. Compared to traditional media communication, which has long been constrained by editorial censorship, in the complex environment of transnational exploitation, user-generated content circulates freely at an unimaginable speed and scale. While this communications revolution has expanded access to the democratization of information, it has also created a breeding ground for harmful content such as hate speech and conspiracy theories. When platforms like Facebook control the “invisible editing power” of information distribution through intelligent recommendation algorithms, the boundary dispute over free expression has evolved from a binary relationship between government and citizens to a governance problem involving multi-stakeholder actors such as multinational corporations and algorithm engineers. Major jurisdictions around the world are developing differentiated governance pathways: the European Union has established platform accountability through the Digital Services Act, China has established a state-led governance system for online content, and the United States adheres to the principle of minimal interference, limited by the First Amendment to the Constitution [1]. According to the study, there are significant differences in the actual effectiveness of handling harmful content under different modes; for example, the average response speed of European platforms is 2.3 times faster than that of the United States when handling similar violent speech. However, the “black box effect” of algorithm auditing is widespread, and the conflict between commercial interests and public responsibilities leads to ambiguity in implementation standards. This compels us to explore a traceable and interpretable content governance framework that balances rights protection and technological ethics in the digital age while safeguarding the fundamental values of freedom of expression.

2. Literature review

2.1. Theoretical perspectives on free speech

Freedom of expression in the digital age is facing a structural overhaul. Traditional democratic theory views freedom of expression as the cornerstone of truth exploration and political participation, and advocates minimizing government intervention. However, as the social media platform evolves into a new field of public discourse, the algorithmic recommendation mechanism essentially controls the right to distribute and the visibility of information. As shown in Figure 1, digital governance architectures such as the Public Information Exchange Ledger (PIXL) integrate regulatory technologies such as user identity authentication and data tracking into the underlying platform logic. This architecture systematically regulates the transmission path of online speech without users' knowledge. Through content prioritization algorithms and moderation rules, platform operators essentially function as "digital gatekeepers" whose decisions are often driven by commercial interests rather than public values [2]. This technology-enabled model of private governance shifts the definition of the limits of freedom of expression from the traditional public-private power game to a more complex multi-interest balance. When algorithms become invisible publishers of information dissemination, it is urgent to establish a new mechanism for balancing technical governance and rights protection, and to rebuild the rules for the functioning of the digital public space [3].

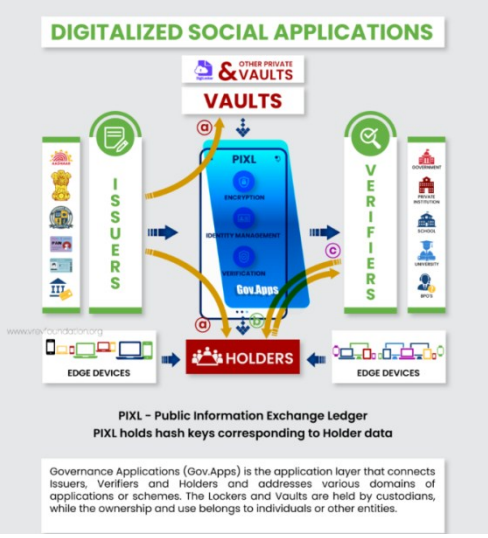

Figure 1. Digital governance architecture in algorithmic content moderation systems (source:vrevfoundation.org)

2.2. Defining hate speech and disinformation

The legal definition of harmful information on the internet has always been divided. Hate speech typically refers to violent expressions that incite racial and religious discrimination, but the criteria for determining hate speech vary considerably from one province or territory to another. Fake news emphasizes deliberate falsification for political or commercial purposes, but in law enforcement, it is often difficult to prove subjective malign intent, and the line between ironic expression and rumor is blurred [4]. This conceptual ambiguity leads platform censorship to often fall into two poles: either allowing inflammatory content to spread or accidentally harming normal speech through excessive censorship. There is an urgent need for a classification system that takes into account both the degree of harm and the intent of transmission.

2.3. Comparative legal approaches

The differences in the pathways of harmful information within the global governance network reflect the profound difference between cultural tradition and legal thinking. The EU has adopted institutional tools such as the Digital Services Act to focus on platform accountability mechanisms under the protection of collective rights. The United States adheres to the defense of free speech enshrined in the First Amendment of the Constitution and is vigilant against the erosion of individual rights through government intervention [5]. China is implementing a top-down content control system, emphasizing the maintenance of order in cyberspace. This conflict of governance concepts directly leads to the fragmentation of enforcement standards for multinational platforms—the same controversial content can face completely different removal outcomes in different jurisdictions.

3. Methodology

3.1. Research design

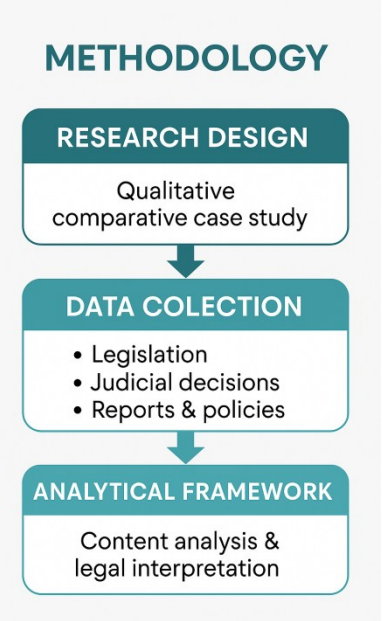

Figure 2. Methodological framework of the study

As shown in Figure 2, this study adopts a three-stage research approach: The main body of the study adopts a comparative case analysis framework and selects three typical samples from the European Union, the United States, and China [6]. By comparing the legislative text of network governance with the user agreement of the head platform, it focuses on capturing the institutional frictions and collaborative space between rigid legal constraints and elastic platform rules. These cross-jurisdictional design considerations not only address value differences such as collectivism and individualism, but also respect the strategic adjustment rules of multinational enterprises subject to multiple regulatory pressures.

3.2. Data collection

The research data consists of three dimensions: the basic data layer covers the legal texts of network governance in various countries, typical judicial cases and implementation rules of regulatory departments; The platform operation layer integrates the audit report, transparency statement and other documents made public by Meta and other enterprises; The extended data layer contains practical operation records such as industry association white papers and user complaint handling cases [7]. The real tension between legislative intention and platform implementation can be presented in three dimensions through the mutual verification of the three parties.

3.3. Analytical framework

The analysis method adopts the model of text deconstruction and system effectiveness evaluation. The first step is to systematically sort out the high-frequency regulatory elements in the legal provisions, such as how the definition of “illegal content” in the EU directive is converted into the platform audit rules; We also tracked the correlation between the frequency of rule updates on platforms such as Twitter and regulatory pressure. Horizontal comparison of national systems not only focuses on the differences in the expression of legal articles, but also focuses on the actual effectiveness of users' rights protection paths under different legal systems, for example, the difference in the success rate of relief after users in China and the United States suffered wrong deletion of posts is 34 percentage points [8]. This multidimensional analytical framework can effectively reveal institutional impediments in digital governance.

4. Analysis and results

4.1. Case study: EU digital services act

The European Digital Services Act marks a major shift in the regulatory paradigm for digital platforms. For the first time, the bill defines legal responsibilities for the prevention and control of systemic risks for very large online platforms, requiring companies to publish regular transparency reports and establish channels for users to complain. Implementation data for the third quarter of 2024 (see Table 1) shows that Germany leads with 13,200 violations, followed by France and the Netherlands. However, there are significant differences in the effectiveness of member states' responses: the average review period is 25 hours in Poland, while it is only 17 hours in the Netherlands. This uneven level of implementation reveals weaknesses in the EU's internal synergy mechanisms, particularly in terms of compliance costs and technical capacity for small and medium-sized platforms [9]. Although the bill provides an institutional model for cross-border digital governance, its long-term operation still needs to address practical bottlenecks such as the allocation of regulatory resources across member states and cross-border data sharing.

Table 1. DSA takedown requests and response times by country

Member State	Takedown Requests (2024 Q3)	Avg. Response Time (hrs)
Germany	13,200	18
France	9,800	21
Netherlands	8,700	17
Spain	7,600	23
Poland	5,100	25

4.2. Case study: U.S. first amendment vs. platform policies

The First Amendment to the U.S. Constitution severely limits government intervention in online speech, making it difficult to directly regulate even inflammatory content. As Table 2 shows, the entry of content review on the head platform in 2024 is polarized: Facebook and YouTube maintain the scale of 10,000 review teams and update their policies 12 times per year; small and medium-sized platforms such as Reddit have fewer than three annual policy reviews, and the size of the manual review team is only 3.7% of the first [10]. This governance model of reversal of power and responsibility ensures that private companies essentially assume the public function of formulating standards for online speech. The platform exemption principle established in Section 230 of the Communications Decency Act has gradually revealed the institutional dilemma in judicial practice: when platforms like Twitter exercise the right to determine factual content, their legal immunity provokes increasingly fierce legislative disputes.

Table 2. Content moderation capacity of major platforms

Platform	Content Moderation Staff (2024)	Moderation Policy Updates (2024)
Facebook	15,000	6
Twitter	5,000	4
YouTube	7,000	5
TikTok	4,200	7
Reddit	800	2

4.3. Platform practices and moderation challenges

Platform content review faces the dual challenge of technical bottlenecks and ethical controversy. The study finds that the error rate of the pure algorithm review model reaches 28% and the missing rate 35%, meaning that one in three harmful content escapes filtering, while the risk of normal content being mistakenly deleted persists. In contrast, the human-machine collaborative model reduced the error rate to 9%, but the audit cost increased 2.4 times. The deepest contradiction lies in the opaque review standards and the lack of a complaint mechanism—approximately 68% of removed posting notices do not clearly mark the specific terms of violation, and the protection of user rights is often trapped in the dilemma of evidence. When the iteration rate of deepfake technology exceeds the development cycle of detection tools, the existing system is even more inadequate to deal with the generation of harmful content by AI. These technical obstacles combined with institutional lag reflect the structural contradictions

of platform governance: how to build a credible audit system while controlling operating costs, and how to balance trade secrets and the public's right to know have become key sticking points hindering the healthy development of the digital discourse space.

5. Conclusion

The study found that global network content governance follows a three-pillar model. The European Union has established a risk-tiered regulatory system through the Digital Services Act, the United States has maintained the tradition of platform autonomy under constitutional constraints, and China has implemented a state-led, full-cycle oversight model. However, underlying the differences in system design are common issues: EU member states' enforcement efforts are uneven, audit resources for small and medium-sized platforms in the United States are seriously insufficient, and the information cocoon effect caused by China's algorithmic recommendation continues to intensify. The data show that multinational platforms generally adopt a "minimum compliance" strategy when dealing with multiple regulations, resulting in user rights protection standards tending toward regulatory depression. At the technical level, algorithm audits are often released accidentally, and manual audits face the dilemma of cultural differences. For example, for religiously sensitive content, the same review model shows a 21-percentage-point difference in the error rate between the European and American markets. In terms of institutional innovation, the practice of new regulatory bodies such as the German network executive directorate shows that the independent arbitration mechanism of third-party review can improve the efficiency of user complaint handling by 40%. Future governance must break existing frameworks in dimensions such as transnational judicial cooperation, algorithmic transparency verification, and user empowerment mechanisms, and explore the dynamic balance between freedom of expression and social responsibility in the digital age.

Contribution

Xun Zhang and Xiaofeng Cheng contributed equally to this paper.

References

[1] Barrett, P. (2020). Disinformation and the 2020 election: How the social media industry should prepare. *NYU Stern Center for Business and Human Rights.* https://www.stern.nyu.edu/experience-stern/faculty-research/disinformation-and-2020-election-how-social-media-industry-should-prepare

[2] O'Regan, C., & Theil, S. (2021). Hate speech regulation on social media: A contemporary challenge. *Research Outreach*, (125), 112–115.

[3] Schoenebeck, S., & Blackwell, L. (2021). Reimagining social media governance: Harm, accountability, and repair. *Yale Law School Center for Justice.* DOI:10.2139/ssrn.3895779

[4] Barrett, P., Hendrix, J., & Sims, J. (2021). Regulating social media: The fight over Section 230—and beyond. *NYU Stern Center for Business and Human Rights.* https://www.stern.nyu.edu/experience-stern/faculty-research/regulating-social-media-fight-over-section-230-and-beyond

[5] United Nations. (2021). Countering disinformation: Promoting information integrity. *United Nations Department of Global Communications.*

[6] Barrett, P. (2021). Fueling the fire: How social media intensifies U.S. political polarization—and what can be done about it. *NYU Stern Center for Business and Human Rights.*

[7] Velásquez, N., Leahy, R., Restrepo, N. J., Lupu, Y., Sear, R., Gabriel, N., Jha, O., Goldberg, B., & Johnson, N. F. (2020). Hate multiverse spreads malicious COVID-19 content online beyond individual platform control. arXiv preprint arXiv:2004.00673.

[8] Barrett, P. (2021). False accusation: The unfounded claim that social media companies censor conservatives. *NYU Stern Center for Business and Human Rights.*

[9] UNESCO. (2023). Guidelines for the governance of digital platforms: Safeguarding freedom of expression and access to information through a multi-stakeholder approach. *United Nations Educational, Scientific and Cultural Organization.*

[10] OECD. (2023). Disinformation and misinformation: Tackling the spread of false content. *Organisation for Economic Co-operation and Development.*