German Law Journal (2025), **26**, pp. 1300–1319
doi:10.1017/glj.2026.10195

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

China's Content Moderation in the Age of Artificial Intelligence

Yilu Zuo

Peking University, China
yilu.zuo@foxmail.com

(Received 30 May 2025; accepted 10 July 2025)

Abstract
This Article discusses China's content moderation in the age of artificial intelligence. It first introduces two long-overlooked features of China's content moderation: the medium-based model and the "No-Dispute" Policy. The former emphasizes that content moderation in China varies based on different media, while the latter argues that China's content moderation is often content-neutral rather than being driven by ideology or having an official stance. The Article then summarizes the three main challenges artificial intelligence presents to content moderation: a shift in structure from the traditional "state v. citizen" dichotomy to the "platform–government–citizen" triangle; a transition in means from human review to algorithm-based and machine-based moderation; and stimulating a reimagination of traditional theories and doctrines of freedom of speech in terms of standards and classification. Finally, the Article takes online violence, one of the most prominent issues in contemporary Chinese content moderation, as a case study to examine specific issues in China's content moderation in the era of artificial intelligence.

Keywords: Content moderation; free speech; artificial intelligence; medium-based model; "No-Dispute"

A. Introduction

To a great extent, content moderation is the flip side of free speech. Much of the attention on China's content moderation is actually a concern for freedom of speech.[1] Unfortunately, attention does not necessarily equal understanding.

In general, there are three tendencies or misunderstandings in China towards freedom of speech and content moderation. The first tendency is a black-and-white dichotomy. People tend to use a black-and-white dichotomy to view China's content moderation.[2] When content moderation is found to exist, it is easy to draw simplistic conclusions such as "China restricts freedom of speech" or "China has no freedom of speech." Under this black-and-white mindset, the determination that content moderation exists often leads to ideological criticism and denial, rather

[1]For more discussion of free speech in China, *see generally* R. Randle Edwards, *Civil and Social Rights: Theory and Practice in Chinese Law Today*, *in* HUMAN RIGHTS IN CONTEMPORARY CHINA 41 (1986); Ellen R. Eliasoph, *Free Speech in China*, 7 YALE J. WORLD PUB. ORD. 287 (1981); Owen M. Fiss, *Two Constitutions*, 11 YALE J. INT'L L. 492 (1986); R. P. Peerenboom, *What's Wrong with Chinese Rights?: Toward a Theory of Rights with Chinese Characteristics*, 6 HARV. HUM. RTS. J. 29 (1993); Roy L. Sturgeon, *China's Homegrown Free-Speech Tradition: Imperial Past and Modern Present. And Post-Modern Future?*, 26 FLA. J. INT'L L. 291 (2014).

[2]*See generally* Mark Jia, *High Theory in Chinese Law*, 103 TEX. L. REV. 382 (2024) (discussing how this kind of dichotomy is common in studies of Chinese law in general).

© The Author(s), 2026. Published by Cambridge University Press on behalf of German Law Journal e.V. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

than a consideration of what Holmes said: that free speech and content moderation are actually more of "a question of proximity and degree."[3]

The second tendency is too abstract and vague. When discussing China's content moderation, it seems as though there is a unified, supreme institution setting and enforcing a uniform set of standards. In other words, traditional wisdom on China's content moderation sometimes lacks nuance, sophistication, and subtlety.

The third tendency is anachronism. The importance of infrastructure and architecture to freedom of speech goes without saying. Although China has long been one of the most advanced markets in new technologies and business models, many perceptions of China's freedom of speech are still based on the printing era of the 19th century.[4] Such views fail to consider the impact of new technologies such as the internet, big data, algorithms, and artificial intelligence on freedom of speech and content moderation.

This Article, therefore, discusses China's content moderation in the age of artificial intelligence. Both artificial intelligence and China's content moderation are highly complex and closely watched topics. Discussing China's content moderation in the AI age is naturally even more complex and challenging. This Article does not claim to provide a comprehensive and definitive answer to this issue; on the contrary, many things, including AI technology itself, are still rapidly evolving and changing. Therefore, this Article merely attempts to provide some hints and outline some possible directions for development.

The Article is divided into three parts. The first part, Section B, is an introduction to China's content moderation, particularly addressing the first two tendencies mentioned above. The first part introduces two long-overlooked features of China's content moderation: the medium-based model and "No-Dispute" Policy. On the one hand, if content moderation is divided into the three questions of "Who moderates," "Moderating whom," and "How to moderate," the medium-based model refers to the fact that China's content moderation actually applies differently based on different media and technologies. On the other hand, "No-Dispute" Policy argues that China's content moderation is not as black-and-white and ideological as commonly thought. In many areas and issues, China's content moderation follows a "No-Dispute" policy. In other words, in many areas, China's content moderation does not uphold an official position; instead, it simply hopes that disputes do not become too polarized and inflamed, leading to serious public opinion or social consequences. Borrowing concepts from the American First-Amendment jurisprudence, "No-Dispute" means that Chinese content moderation is sometimes not viewpoint-based, but rather viewpoint-neutral or keywords-based.[5]

The second part, Section C, addresses the third tendency, anachronism, which attempts to bring our imagination of freedom of speech and content moderation from the 19th century into the present day. Today, the transformation of freedom of speech and content moderation mainly manifests in three points: First, in terms of structure, content moderation has shifted from the original "state v. citizen" dichotomy to the "platform–government–citizen" triangle. Second, in terms of means, content moderation has shifted from human review to algorithm-based and machine-based moderation. Third, in terms of standards and classification, classical theories and doctrines of freedom of speech also face great challenges.

Finally, the third part, Section D, takes one of the most prominent issues in today's Chinese content moderation—online violence—as a case study to discuss some specific issues of China's content moderation in the age of artificial intelligence.

[3]*See* Schenck v. United States, 249 U.S. 47, 52 (1919).

[4]*See generally* Jerome A. Barron, *Access to the Press: A New First Amendment Right*, 80 HARV. L. REV. 1641 (1967) (discussing the 19th-century imagining of free speech).

[5]Frederick Schauer, *Categories and the First Amendment: A Play in Three Acts*, 34 VAND. L. REV. 265, 284–85 (1981).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

B. Two Overlooked Features of Content Moderation in China

This section will provide a simple introduction to China's content moderation. More specifically, it will focus on two long-overlooked characteristics of China's content moderation: the medium-based model and the No-Dispute policy. Hopefully, through this introduction, readers can reduce some overly ideologized and oversimplified understandings of China's content moderation, and recognize more nuance, sophistication, and subtlety.

First, take a look at the typical regulations on content in Chinese law. Article 12 of the Cybersecurity Law states:

> It is not allowed to use the Internet to engage in activities that harm national security, honor, and interests, incite subversion of state power, overthrow the socialist system, incite separatism, undermine national unity, promote terrorism, extremism, ethnic hatred, ethnic discrimination, spread violence, obscene and pornographic information, fabricate and disseminate false information to disrupt economic and social order, as well as activities that infringe on the reputation, privacy, intellectual property rights, and other legitimate rights and interests of others.[6]

Similar provisions can also be seen in Article 15 of the Measures for the Administration of Internet Information Services,[7] Article 6 of the Regulations on the Moderation of Internet Content Ecology,[8] Article 25 of the Regulations on Publishing Moderation,[9] and Article 32 of the Regulations on Radio and Television Moderation.[10] Such provisions are often referred to as "X no-nos" (X不准, *X Bu Zhun*). Simply based on the wording, the provisions on "X no-nos" in various legal documents are actually quite similar, and it is highly probable that two conclusions will be reached: First, China has content moderation or restricts speech; second, China moderates certain types of content more strictly. In the mindset of black-and-white thinking, when the conclusion regarding the presence of content moderation is "yes," it often leads to ideologized criticism.

But this mindset actually provides limited help for us to understand content moderation or promote free speech in China. As mentioned earlier, we may need a more nuanced and sophisticated analysis. For example, content moderation can be further subdivided into three sub-issues: first, who moderates?; second, moderating whom?; and third, how to moderate? For the first two questions, traditional studies tend to, on one hand, use terms like "Chinese government" or "propaganda department" to refer to all content-moderation entities; on the other hand, such studies treat content of different media as abstract targets. For the third question, the traditional view tends to use Western concepts and theories to imagine and evaluate China's practices. With

[6]Zhōnghuá rénmín gònghéguó wǎngluò ānquán fǎ (中华人民共和国网络安全法) [Cybersecurity Law] (promulgated by the Standing Comm. Nat'l People's Cong., Nov. 7, 2016, effective June 1, 2017), https://flk.npc.gov.cn/detail?id=2c909fdd678bf17901678bf8276f093d&fileId=&type=&title=%E4%B8%AD%E5%8D%8E%E4%BA%BA%E6%B0%91%E5%85%B1%E5%92%8C%E5%9B%BD%E7%BD%91%E7%BB%9C%E5%AE%89%E5%85%A8%E6%B3%95.

[7]Hùliánwǎng xìnxī fúwù guǎnlǐ bànfǎ (互联网信息服务管理办法) [Measures for the Administration of Internet Information Services] (promulgated by State Council, Jan. 8, 2011, effective Jan. 8, 2011), https://flk.npc.gov.cn/detail?id=ff8080816f3cbb3c016f4118e47616f1&fileId=&type=&title=%E4%BA%92%E8%81%94%E7%BD%91%E4%BF%A1%E6%81%AF%E6%9C%8D%E5%8A%A1%E7%AE%A1%E7%90%86%E5%8A%9E%E6%B3%95.

[8]Wǎngluò xìnxī nèiróng shēngtài zhìlǐ guīdìng (网络信息内容生态治理规定) [Regulations on the Moderation of Internet Content Ecology] (promulgated by Cyberspace Admin., Dec. 15, 2019, effective Mar. 1, 2020), Off. Cent. Cyberspace Affs. Comm'n, Dec. 20, 2019, https://www.cac.gov.cn/2019-12/20/c_1578375159509309.htm.

[9]Chūbǎn guǎnlǐ tiáolì (出版管理條例) [Regulation on the Administration of Publication] (promulgated by State Council, Dec. 6, 2024, effective Jan. 2, 2025), https://flk.npc.gov.cn/detail?id=ff8081819667377001 96a8c35f863256&fileId=&type=&title=%E5%87%BA%E7%89%88%E7%AE%A1%E7%90%86%E6%9D%A1%E4%BE%8B.

[10]Guǎngbò diànshì guǎnlǐ tiáolì (广播电视管理条例)[Regulations on Radio and Television Moderation] (promulgated by State Council, Aug. 11, 1997, effective Sep. 1, 1997; rev'd by the State Council, Nov. 29, 2020), https://flk.npc.gov.cn/detail?id=ff808081777d07c50177b8e2bc3f391b&fileId=&type=&title=%E5%B9%BF%E6%92%AD%E7%94%B5%E8%A7%86%E7%AE%A1%E7%90%86%E6%9D%A1%E4%BE%8B.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

the help of the medium-based model and No-Dispute policy, we may discover that content moderation in China actually has its own practices and logics, both of which are different from those in the West.

I. A Medium-Based Model

According to the medium-based model, China's content moderation adopts different moderation methods and standards based on the attributes of different mediums.[11] The medium-based model is mainly based on two attributes of the medium: technical attributes and social attributes. The former refers to the technical characteristics of a certain medium, while the latter includes the nature of the medium in terms of culture, ideology, and other aspects in a specific society. China's content moderation has never been one-size-fits-all, but is customized based on the technical and social attributes of different mediums. Which medium should be moderated (or moderated to higher standards), which department should be in charge, and how and by what standards to moderate, the answers to these questions are all different. Moreover, medium-based content moderation is dynamic rather than static. When the attributes of the medium change, "who moderates," "moderates whom," and "how to moderate" may also change accordingly.

To clarify, this Article uses "medium" in a broad sense—not to the extent of "all is medium," but indeed the definition includes the vast majority of intermediaries, forms, and places that carry content and information. The mediums discussed in this Article can be broadly divided into five categories: first, live events; second, print media; third, radio and television; fourth, films; and fifth, the Internet.

Though they may not have noticed it, most people who live in China encounter the medium-based model daily. For example, on September 14, 2020, the Ministry of Culture and Tourism issued the "Notice on Deepening the Reform of Delegating Power, Strengthening Regulation, and Improving Services to Promote the Prosperous Development of the Performance Market," which mentioned the need to strengthen supervision over "key performance types" such as electronic music shows, rap programs, immersive performance activities, small-theater stand-up comedy and crosstalk, and avant-garde and experimental dramas.[12] Likewise, in 2022, the Shanghai Municipal Administration of Culture and Tourism, recognizing that "escape rooms" and "scripted murder mystery games" (剧本杀, *Ju Ben Sha*) have gained popularity among consumers, especially the younger demographic, due to their novel entertainment formats and strong social attributes, announced the "Interim Provisions on the Content moderation of Escape Rooms and Scripted Murder Mystery Games in Shanghai."[13] As a result, Shanghai became the first city in the country to include "scripted murder mystery games" in its content-moderation system.

Why did the regulatory authorities choose to regulate or enhance the regulation of the aforementioned media and formats rather than others? Why did they choose this particular point in time to implement regulations? More specifically, why did they introduce new content-moderation regulations specifically targeting scripted murder mystery games, electronic music,

[11]*See* Yilu Zuo, *Juti Meijie Juti Guanli – Zhongguo Meijie Neirong Guanli Moshi Chutan* (具体媒介具体管理——中国媒介内容管理模式初探) [*Medium-by-Medium Regulation: the "Medium-Based" Paradigm of China's Content Moderation*], *in* 2 Kaifang Shidai (开放时代), Open Times, 191–206 (2023) (discussing moderation methods and different mediums).

[12]Wénhuà hé lǚyóu bù guānyú shēnhuà "fàng guǎn fú" gǎigé cùjìn yǎnchū shìchǎng fánróng fāzhǎn de tōngzhī (文化和旅游部关于深化"放管服"改革促进演出市场繁荣发展的通知) [Notice on Deepening the Reform of Delegating Power, Strengthening Regulation and Improving Services to Promote the Prosperous Development of the Performance Market] (promulgated by The Ministry of Culture & Tourism, Sep. 14, 2020), https://zwgk.mct.gov.cn/zfxxgkml/scgl/202012/t20201206_918203.html.

[13]Shànghǎi shì mìshì jùběn shā nèiróng guǎnlǐ zhàn háng guīdìng (上海市密室剧本杀内容管理暂行规定) [Shanghai Interim Regulations on the Management of Escape Room Script-Killing Content] (promulgated by the Shanghai Mun. Admin. Culture & Tourism, Jan. 13, 2022, effective Jan. 13, 2022), https://whlyj.sh.gov.cn/jqxxgk/20220113/bc6633f27d8e4b3bac3394cc2a6d8c4d.html.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

rap, small-theater stand-up comedy, and crosstalk—and not "Werewolves of Miller's Hollow" (*Lang Ren Sha*), opera, pop music, Peking Opera, traditional drama, large-scale commercial stand-up comedy and crosstalk? Moreover, why are there three different standards for works like "The Joy of Life" (*Qing Yu Nian*) when they are serialized on websites like Qidian Chinese Literature and Zongheng Chinese Literature, published as novels, and adapted into TV dramas? For TV dramas, why are there differences in the regulation of online dramas and terrestrial dramas in terms of themes such as time travel, fantasy, pursue immortality through cultivation, and criminal investigation, as well as whether characters can smoke or cohabit before marriage? For games, why are there different standards between those with storylines and those without? Even for the same song, why can certain lyrics be normally included in physical albums and music apps, but some words or phrases have to be replaced when performed on television?

The above examples demonstrate that China's content moderation is medium-based. As mentioned earlier, this content-moderation model is mainly based on two attributes of the medium: technical attributes and social attributes. Technical attributes are the essence or technical characteristics of a particular medium. Harold Innis argues that different media have different biases—for example, papyrus and paper are convenient for use and transportation but are not for preservation, so they are more biased towards space; stone carvings and clay tablets are good for long-term preservation, but are not convenient for use and transportation, so they are more biased towards time.[14] What Innis calls "bias" is actually the technical attributes of different media.

Technical attributes are relatively easy to understand. There are two reasons for this: first, compared to social attributes, the technical attributes of a particular medium are often relatively fixed. Whether in China, the United States, or South Africa, the technical attributes of broadcasting are basically the same, but the social attributes may differ. Second, it is a common practice in many countries to regulate based on the technical attributes of different media. The most typical example in this regard actually comes from the United States. In *Red Lion v. FCC*, the U.S. Supreme Court ruled that because broadcasting is subject to frequency scarcity, it is different from newspapers—as a result, different regulatory standards could apply, and the government was free to regulate broadcasting content more strictly.[15] And in *FCC v. Pacifica Foundation*, the U.S. Supreme Court—based on other technical attributes of broadcasting, namely invasiveness and pervasiveness—held that the government could regulate obscene content on radio.[16] Whether it is frequency scarcity, invasiveness or pervasiveness, they are all technical attributes of broadcasting; it is also based on these technical attributes of broadcasting that the U.S. Supreme Court decided to adopt different content moderation standards for broadcasting as opposed to newspapers.

Technical attributes can explain some of the examples mentioned earlier. Jack Balkin argues that content moderation should be based on filterability—whether a medium has an effective information filtering mechanism.[17] A high level of filterability in a medium means that recipients of content can more easily select, organize, and shield information; conversely, if recipients find it difficult to select, organize, and shield certain information, the medium's filterability is lower.[18] Content moderation based on the technical attribute of filterability means reducing regulation on media with high filterability, while strengthening content moderation on media with low filterability. From the perspective of filterability—compared, for example, to murder-mystery games and traditional plays—the need for enhanced regulation in scripted murder mystery games and immersive theater is due to factors such as interactivity and non-player characters (NPCs) which significantly reduce the filterability of such performances. Similarly, although the audience

[14]*See* Harold Innis, The Bias of Communication 3–32 (1951).

[15]Red Lion Broad. Co. v. Fed. Commc'ns Comm'n, 395 U.S. 367, 400 (1969).

[16]Fed. Commc'ns Comm'n v. Pacifica Found., 438 U.S. 726, 745–51 (1978).

[17]Jack M. Balkin, *Media Filters, the V-Chip, and the Foundations of Broadcast Regulation*, 45 Duke L.J. 1131, 1141–44 (1996).

[18]*Id.* at 1141–44.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

size at small theater performances of crosstalk and stand-up comedy is much smaller than that in venues like sports stadiums, in small theaters actors interact more frequently with the audience and the intimacy of the venue may lead actors to consciously or unconsciously relax certain "boundaries," resulting in decreased filterability. Regulators' focus on rap music above other music genres, apart from certain subcultural factors, is also due to filterability. Indeed, almost all forms of performance involve an element of improvisation, but rap's emphasis on freestyle makes it easier for uncontrollable content to arise in rap music.

The analysis of technical attributes above focuses more on "who to moderate" and "how to moderate," and less on "who moderates." However, the emergence of new media and new technological attributes will also change the question of "who moderates." In the United States, the establishment of the Federal Communications Commission (FCC) as a regulatory body coincided with the emergence of broadcasting. On the question of "who moderates," what makes China unique is that it not only regulates based on different technical attributes but also establishes different institutions specifically for the content moderation of certain media. For a considerable period, China's institutions involved in content moderation have included the State Administration of Radio and Television, the National Film Administration, the General Administration of Press and Publication, the State Council Information Office, the Cyberspace Administration of China, the Ministry of Culture, the Ministry of Industry and Information Technology, and the Central Propaganda Department, among others. Printed publications were once under the jurisdiction of the General Administration of Press and Publication, television programs were moderated by the State Administration of Radio and Television, movies were overseen by the National Film Administration, the internet was regulated by the Cyberspace Administration of China, offline performances were under the moderation of the Ministry of Culture, and so on. There have been overlaps and conflicts, such as the debate between the Ministry of Culture and the General Administration of Press and Publication regarding the license of video games in the past, as well as the widely mentioned "Nine Dragons Governing the Waters" in discussions of internet regulation. To a certain extent, newspapers, magazines, books, radio and television, movies, the internet, live performances, and games each have one or even several corresponding institutions for moderation. The fact that different departments moderate different media is a prime example of China's medium-based content moderation model. Yet as will be analyzed later, changes in the question of "who moderates" are often closely linked to the social attributes of the medium.

Compared to technological attributes, social attributes are more subtle and have more "Chinese characteristics." Social attributes are a relatively broad concept, including but not limited to the function, popularity, and ideological importance of a certain medium within a particular society. As mentioned before, while the technological attributes of a medium may be largely similar across countries, the social attributes can differ greatly. If technological attributes can explain why different content moderation approaches are taken for different media, social attributes can answer why many countries adopt different approaches for the same medium and why the content moderation of the same medium in the same country changes over time. As one might imagine, social attributes play an important role in content moderation in China. To a certain extent, one could even say that social attributes are the key to understanding content moderation in China.

Take crosstalk as an example. Technological attributes can explain why there needs to be stricter regulation of small-theater crosstalk performances rather than large commercial shows, but social attributes can help us understand why there is now a need to strengthen the regulation of crosstalk in general. Although the technological attributes of crosstalk have not changed much, its social attributes have changed in recent years. From the 1990s to the early 2000s, crosstalk was extremely marginal. Under this social attribute, even if inappropriate content appeared in crosstalk performances at that time, its impact and harm would be very limited because no one was watching. However, with the revival of crosstalk by Guo Degang and Deyun Society in the early 2000s, crosstalk has gone from being ignored to becoming one of the most popular art forms

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

today. In addition to its becoming extremely popular, another important variable in the changing social attributes of crosstalk is its younger audience demographic. When the crosstalk audience shifted from relatively older, traditional fans to young people proficient in using the internet and deeply influenced by "fandom culture," strengthening content moderation for crosstalk became inevitable. The underlying logic is the assumption that young people are more "susceptible"—that is, they are more likely to be influenced and may react more strongly or become more polarized after being influenced. For the same reason, social attributes can also explain why content moderation needs to be strengthened for genres like hip-hop, electronic music, stand-up comedy, immersive avant-garde theater, and scripted murder-mystery games, because they are also deeply loved by young people.

Social attributes are not only reflected in "moderate whom" and "how to moderate," but also in "who moderates." Changing "who moderates" based on changes in the social attributes of a medium is a very distinct characteristic of content moderation in China. The most typical example of this is film. The department responsible for regulating the content of Chinese films has changed multiple times throughout history, and as the following discussion demonstrates, almost every change was related to shifts in the social attributes of film.

After the founding of the People's Republic of China, films were initially managed by the Ministry of Culture, with the main task being to take over and rebuild the film enterprises and institutions from the former KMT government to establish a film industry for the people. Then in January 1986, the 6th National People's Congress Standing Committee reviewed and approved the "Decision to Change the Ministry of Radio and Television to the Ministry of Radio, Film and Television." This move integrated the management of radio, television, and film, with the Film Bureau being transferred from the Ministry of Culture to the new Ministry of Radio, Film and Television.[19] This change was mainly because radio and television were developing rapidly at the time, so it was hoped that the well-developed radio and television could boost the development of the film industry. However, although radio and television continued to develop rapidly, the film industry fell into a slump in the late 1980s and early 1990s, essentially exiting people's public cultural life.

Starting in 1994, however, China began importing some Hollywood films each year, and the film market began to improve. From 2000 onwards, the Chinese film market and Chinese films ushered in the so-called "fourth boom" or "golden 20 years."[20] By 2018, China's domestic box office reached 60.976 billion, accounting for about 21.34% of the global box office, with a total of 60,079 screens and 1.716 billion urban moviegoers.[21] A more intuitive sense is that audiences (especially young audiences) not only returned to cinemas, but film-related topics also re-entered public discussion. It was against this backdrop that in the 2018 "Plan for Deepening the Reform of Party and State Institutions" issued by the CPC Central Committee, the film management responsibilities of the State Administration of Press, Publication, Radio, Film and Television were transferred to the Central Propaganda Department, which took on the role of the State Film Administration. Film management shifted from the State Administration to the Central Propaganda Department. The document particularly emphasized that this change was "to better play the special and important role of film in propaganda, ideology, and cultural entertainment, and to develop and prosper the film industry."[22] The "special and important role of film in

[19]Zhao Yuming, Zhongguo Guangbo Dianshi Tongshi (中国广播电视通史) [A General History of Chinese Radio and Television] 389 (2011).

[20]Yin Hong & Liang Junjian, *The "Fourth Upsurge" in the Development of Chinese Cinema — On Film Creation Since the New Century*, *in* Modern Communication 70, 71 (2021).

[21]While writing this Article, the film "Ne Zha: The Demon Boy II" achieved a box office of 2.26 billion US dollars during the Spring Festival film schedule in 2025, ranking fifth on the global box office chart of film history. *See Ne Zha 2*, Box Office Mojo, https://www.boxofficemojo.com/release/rl1613856769/.

[22]Zhōnggòng zhōngyāng guānyú shēnhuà dǎng hé guójiā jīgòu gǎigé de juédìng (中共中央关于深化党和国家机构改革的决定) [Decision of the CPC Central Committee on Deepening the Reform of the Party and State Institutions] (promulgated by the Cent. Comm. Communist Party China, Feb. 28, 2018, effective Feb. 28, 2018).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

propaganda, ideology, and cultural entertainment" mentioned here actually refers to the social attributes of film.

II. No-Dispute Policy

Another frequently overlooked characteristic of China's content moderation is the "No-Dispute" policy. This concept was created by Deng Xiaoping. In "Excerpts from Talks Given in Wuchang, Shenzhen, Zhuhai and Shanghai," which is the very last essay of the Selected Works of Deng Xiaoping, Deng emphasizes: "No-Dispute is my invention, so as to have more time for action. Once disputes begin, they complicate matters and waste a lot of time. As a result, nothing is accomplished. No dispute; try bold experiments and blaze new trails."[23]

The "No-Dispute" policy has its historical background, with its core focusing on depoliticization and de-ideologization. Deng proposed this policy in the early stages of China's reform and opening up. The Cultural Revolution caused tremendous harm to China, with one of the roots of the Cultural Revolution being the emphasis on "class struggle." Therefore, "No-Dispute" first and foremost means shifting the focus of the party and the state from class struggle to economic development. Additionally, in the 1980s, many voices believed that reform and market reforms were not in line with traditional socialist orthodoxy and were "taking the capitalist road." In this sense, "No-Dispute" also calls for refraining from engaging in ideological debates about whether to "take the capitalist road" or "take the socialist road," and to allow more tolerance and time for reforms.[24]

Just like many legacies of Deng Xiaoping continue to profoundly influence China today, "No-Dispute" provides us with a key to understanding China's content moderation today too. Simply put, many foreign observers may perceive China's content moderation as highly ideological, with departments responsible for censorship and content moderation strictly enforcing party and government policies. Only speech that aligns with the party and government's position can be published, while speech that does not align with their position cannot be published.

Contrary to the above understanding, this Article argues that China's content moderation is not as black-and-white and ideological. In many aspects, China's content moderation follows the "No-Dispute" policy. In other words, in many fields, China's content moderation does not uphold an orthodox position; instead, it simply hopes that disputes do not become overly heated and polarized, leading to serious public opinion or social consequences. Using concepts from United States First Amendment jurisprudence, "No-Dispute" means that China's content moderation is sometimes keywords-based instead of viewpoint-based.[25] Discussions on certain topics may indeed be subject to stricter moderation, but it is not about the suppression of voices opposing the official position by content moderation departments and platforms. Rather, the authorities do not have a stance on the issue and simply hope that both sides do not exacerbate the conflict—or, in order to implement "No-Dispute," all voices are restricted.

One of the most typical examples in this regard is the practice of "prohibiting the creation of extreme gender opposition." Like many countries, gender equality is one of the most discussed topics on the Chinese internet, and it is also very controversial. Controversies are particularly prominent on issues related to employment, childbirth, and "dowry." According to stereotypes, whether based on Confucian tradition or an ideology different from the West, China's official stance tends to be more in favor of male power and patriarchal dominance, or less friendly to feminism. But this is not the case. Just as in many issues, China's content moderation enforces a

[23]Deng Xiaoping, *Excerpts from Talks Given in Wuchang, Shenzhen, Zhuhai and Shanghai*, *in* Selected Works of Deng Xiaoping (1982–1992), at 362 (1994).

[24]*See* Yilu Zuo, *Ideological and Cultural Policy as Speech Policy: An Overlooked Aspect of Free Speech in China (1940s–2000s)*, 11 Asian J.L. & Soc'y 382, 395 (2024) (discussing the capitalist and socialist roads further).

[25]*See* R.A.V. v. City of St. Paul, 505 U.S. 377 (1992).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

"No-Dispute" policy on gender issues. Platforms like Weibo and Douyin have community guidelines prohibiting the creation of gender opposition.[26] The Cyberspace Administration's campaign-style governance on content moderation, the "Clear and Bright Action," has also specifically targeted the creation of gender opposition. Just as the name "prohibition of creating gender opposition" reflects, the regulatory authorities neither favor feminism nor uphold male power and patriarchal dominance. They focus on a viewpoint-neutral consequence—preventing extreme gender opposition. "No-Dispute" manifests here as, regardless of what stance you support, as long as you do not incite "extreme gender opposition," you are allowed; but if your voice may incite "extreme gender opposition," then regardless of your stance, restrictions will be imposed.

Although in terms of equality, "no-dispute" is actually relatively fair because it is not viewpoint-based and does not favor either side in the debate. But in terms of freedom, "no-dispute" can indeed limit more speech. Of course, this Article is not defending this policy, but simply pointing out that China's content moderation has this characteristic.

As mentioned earlier, the medium-based model and "No-Dispute" are two long-overlooked characteristics of content moderation in China. To understand China's content moderation in the age of artificial intelligence, one must first understand content moderation in China. And these two characteristics are key to understanding content moderation in China. However, these two features have existed since the era of print media, rather than being brought about by new technologies and artificial intelligence. Technologies like the internet and artificial intelligence have brought about new challenges for content moderation worldwide. The comparative complexity of content moderation in China arises because China is facing not only the challenges commonly faced by other countries but also unique difficulties. In the case of the latter challenge, many are related to the medium-based model and "no-dispute." Therefore, the next section will discuss both the universal challenges and difficulties in content moderation and examine the unique problems in China.

C. Content Moderation in the Age of AI

This section attempts to correct the anachronisms in free speech and content moderation, updating our imagination of free speech and content moderation from the 19th century to today. To discuss the challenges that content moderation faces in the age of artificial intelligence, it may be necessary to first clarify that the various challenges confronting content moderation today may not solely be brought about by artificial intelligence *per se*. Instead, we should consider this issue from a more macro and long-term perspective. In other words, many of the problems faced by content moderation today may have started emerging since the birth of the internet or even the rise of mass media such as broadcasting. This is a much longer process, and artificial intelligence is just the latest development in this series of technological advancements.

Often, content moderation and freedom of speech are two sides of the same coin. Many scholars have pointed out that classical thinking on freedom of speech actually originated from the understanding and imagination of print media in the 19th century.[27] Owen Fiss summarizes the classical imagination of freedom of speech as the "the street corner speaker" paradigm.[28] The emergence of radio and broadcast television was the first major challenge to this paradigm. The previously mentioned *Red Lion* decision is a typical representation of this challenge. Fiss believes that the emergence of radio and broadcasting television means that all speakers have moved away

[26]*See Community Guidelines*, WEIBO, at art. 36, https://service.account.weibo.com/h5/roles/gongyue?ua=HONOR-ELZ-AN00__weibo__12.8.1__android__android12 (last visited Oct. 27, 2025).

[27]*See generally* Jerome Barron, *Access to the Press — A New First Amendment Right*, 80 HARV. L. REV. 1641 (1967).

[28]Owen Fiss, *In Search of a New Paradigm*, 104 YALE L. J. SYMP. 1613, 1614 (1995).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

from the street corner and towards CBS, necessitating a paradigm shift in freedom of speech and content moderation.[29]

Compared to radio and television, the emergence of the internet undoubtedly posed a much greater challenge. Cyberlaw scholars are familiar with the debate on the "law of horses," which discusses whether the internet poses a comprehensive and fundamental challenge to legal theories.[30] Scholars such as Larry Lessig believe that the changes and impacts brought about by the internet are significant and fundamental enough.[31] In the area of freedom of speech and content moderation, the internet has once again necessitated a "search for a new paradigm."[32] Following the internet, we have seen the emergence of big data, algorithms, and artificial intelligence, which are topics of discussion today. Therefore, the term used in this Article is content moderation in the AI era, emphasizing that our focus is on the challenges faced by content moderation today, rather than just the challenges brought about by artificial intelligence technology *per se*.

Overall, this Article holds that content moderation faces the below challenges in the era of artificial intelligence.

I. Structure: From "State v. Citizen" Dichotomy to "Platform–Government–Citizen" Triangle

In terms of structure, content moderation is transitioning from the dichotomy of "state v. citizen" to the triangle of "platform–government–citizen."[33] Using the concept of the "three-body problem" coined by the renowned Chinese Sci-Fi writer Liu Cixin, content moderation has now become a complex issue involving all three entities.[34] In this shift from dichotomy to a triangular relationship, the most prominent factor is the role and power of the platforms.

For a long time, the demarcation between private and public has been one of the basic premises of classical legal and political theory. The dichotomy of "state v. citizen" in freedom of speech and content moderation is based on this idea. Under such a conception, as Fiss argues in *The Irony of Free Speech*, because the government is mainly responsible for content moderation and speech regulation, the government is often seen as the "greatest enemy" of free speech.[35] However, the rise of the internet and platforms has completely changed this dichotomy. People living today are certainly no strangers to this change. The influence and power of platforms on how people express and disseminate information today may be no less than—and even far greater than—that of the government. To continue borrowing the concept of "the street corner speaker," if someone wants to participate in public discourse today, the street corner does not provide them with the surest avenue; they instead must go through platforms. Whether it was the previous de-platforming of Trump by social media platforms or the controversy surrounding the US Congress' mandatory divestment of TikTok in 2024, both prove the importance of platforms from different angles. The emergence of the very concepts of content moderation and governance also indirectly prove that today, it is not only the government's censorship and regulation, but multiple entities, including platforms, that participate in moderation.

The impact of the "platform–government–citizen" triangle on content moderation is reflected in the following ways:

In the context of the triangular relationship, when people discuss content moderation, platforms have replaced the government as the main target. As Kate Klonick aruges, platforms have become the new governor.[36] In China, when discussing content moderation, a frequently

[29]*Id.*

[30]Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 503 (1999).

[31]*Id.*

[32]*See generally* Fiss, *supra* note 28.

[33]Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2011 (2018).

[34]*See generally* CIXIN LIU, THE THREE-BODY PROBLEM (Ken Liu trans., Tor Books 2014) (2008).

[35]OWEN M. FISS, THE IRONY OF FREE SPEECH 27–49 (1996).

[36]*See generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

used concept is "principal responsibility" (主体责任, *Zhu Ti Ze Ren*).[37] Principal responsibility refers to the active obligations and potential adverse consequences that platforms, as market operators and organizers and managers of related activities, should actively assume during their business operations and service provision process, as well as the corresponding adverse consequences that may result from failing to fulfill obligations. Put simply, principal responsibility includes at least two basic requirements: First, when it comes to content moderation, platforms should proactively take responsibility and actively moderate content; second, if something happens, the government will primarily hold the platform—not the users—accountable.

Of course, on the one hand, some people may argue that "principal responsibility" is the government's way of avoiding responsibility. On the other hand, however, principal responsibility also makes a lot of sense. Considering the role of platforms today and the technological capabilities they possess, if one must choose a party to bear primary responsibility, the platform is the most rational choice. The creation and widespread acceptance and use of concepts such as "principal responsibility" or "new governor" largely demonstrate that platforms have indeed become the primary responsible party in content moderation.

How will AIGC (Artificial Intelligence Generated Content) affect the triangle and platforms? From the perspective of weakening monopolies, AIGC may challenge existing big tech companies and bring about some reshuffling possibilities. Some new platforms may emerge as leaders with more advanced technology and business models, while older platforms may fall behind. This is also why existing tech giants and platforms are heavily investing in AI. Although some new tech giants may replace old platforms, the monopolies held by new platforms may be more difficult to shake. The reasons for this are manifold. First, AI may completely revolutionize how people express and create. While platforms have played a significant role in dissemination in the past, with the emergence of AIGC, people may increasingly rely on platforms for creation and expression as well. Second, with platforms collecting more data and gaining more precise algorithms, they can further enhance user experience and stickiness, forming stronger competitive barriers. In other words, in the age of artificial intelligence, the monopolistic and controlling position of platforms respecting content creation and expression may be even more formidable.

The triangle will make "the irony of free speech" more prominent. When discussing the irony of free speech, Fiss mainly focuses on mass media such as broadcast television. Fiss believes that in order to combat mass media, citizens sometimes need to rely on the government—the former greatest enemy of free speech—to protect their own freedom of speech.[38] In the age of artificial intelligence, platforms have more power than traditional mass media, but the dynamics and logic described by Fiss still hold true. This means that, under the triangle of "platform–government–citizen," citizens sometimes need to also rely on the power of the government to resist the platform. More specifically, under the triangle, on one hand, as highlighted by the concept of "principal responsibility," the government will push platforms to the forefront to take on the role of "bad guys" responsible for content moderation. Concepts such as collateral censorship and public/private cooperation/cooptation, as coined by Jack Balkin, illustrate this point.[39] On the other hand, platforms also have their own agenda and may engage in content moderation based on their own interests and ideologies—what Balkin termed "private governance."[40] In these cases, government intervention is necessary to restrain the platform.

[37]*See generally* Dai Xin, *Platform Responsibility and Social Trust*, *in* 2 Science of Law 82 (2023).
[38]Fiss, *supra* note 35, at 27–49.
[39]Balkin, supra note 33, at 2015–21.
[40]*Id.*

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

II. Means: From Human Review to Algorithm and Machine-Based Moderation

In terms of tools or methods, compared to the print and broadcast eras, the internet age generates massive amounts of content every moment. AI and big data will undoubtedly further exacerbate this trend, requiring today's content moderation to utilize new technologies such as machines and algorithms, as opposed to past methods of reviewing one case at a time by human. As Evelyn Douek has noted, content moderation thus needs to move towards proportionality and probability.[41]

The traditional content moderation of legacy media was essentially one case at a time, and the U.S. Supreme Court may also only need to decide a few cases related to free speech in a year. Traditional content review and moderation was primarily human-based, with sufficient time for detailed and unhurried assessments. Under this premise, people had reasonable expectations that content moderation or censorship must be relatively precise. However, in the internet age platforms need to manage massive amounts of content every moment. It is claimed that Facebook averages around 1.1 million content reviews per day.[42] Content moderation of this magnitude cannot rely on human efforts alone, and must introduce technologies such as artificial intelligence and algorithms, but this will inevitably lead to errors. Therefore, content moderation based on probability first implies accepting the possibility of mistakes. The emergence of AIGC technologies will make expression and creation easier, requiring platforms to review more content every minute; this also forces platforms to introduce new technologies and methods like AI for content moderation. However, while improving efficiency, these new technologies and methods will inevitably be less accurate and cause some collateral damage. Those still accustomed to the relatively precise old-school regulation must learn to accept the efficient but rough new-school regulation.[43]

Meanwhile, accepting proportionality implies changing the understanding of free speech absolutism. Dworkin has argued for "rights as trumps,"[44] but in today's heterogeneous society free speech is no longer a trump card. Instead, when moderating content, one should fully recognize the plurality of values and the relativity of rights, then balance free speech with other purposes as well as the costs and benefits of regulatory measures according to the principle of proportionality in specific contexts. Furthermore, considering the rise of the proportionality principle globally in past decades, adopting this approach may also increase the legitimacy of content moderation from a cultural and psychological perspective. For example, Facebook's Oversight Board adopted the proportionality principle in its decision on Trump's appeal.[45] The spreading adoption of the proportionality principle is a trend worth continued attention.[46]

III. Standard and Classification: Challenges to the Classical Free Speech Theory and Doctrine

In terms of standards and classification, issues and controversies like fake news and hate speech are causing some of the doctrines and theories of classical freedom of expression to waver.

First, it may be necessary to allow content-based content moderation. Traditionally, the prohibition of "content-based" regulation has been considered a cornerstone of freedom of speech. Regulations that do not relate to the content of speech, but only involve time, place, and manner, are often allowed. Regulations that involve content are often seen as unconstitutional. For example, it may be permissible to require no speeches using amplifiers in the town square after 8

[41]Evelyn Douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 759 (2021).

[42]*Id.* at 791.

[43]*See generally* Jamal Greene, *The Supreme Court 2017 Term — Foreword: Rights as Trumps?*, 132 HARV. L. REV. 28 (2018).

[44]*See generally* RONALD DWORKIN, TAKING RIGHTS SERIOUSLY (1977); Greene, *supra* note 43, at 32–38.

[45]*Former President Trump's Suspension*, OVERSIGHT BOARD (May 5, 2021), https://www.oversightboard.com/decision/FB-691QAMHJ/.

[46]*See generally* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296 (2014).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

PM, but it cannot be required to prohibit political speech or support for a certain political view in the town square. The rationale behind this prohibition is that once regulators are allowed to regulate based on the content of speech, they may suppress speech they do not like or promote speech they favor, which would interfere with the "marketplace of ideas" and freedom of speech. Under the principle of prohibiting "content-based" regulation, public discourse has gradually formed a formality and relativism that does not inquire about the content of speech. Freedom of speech does not care about the quality of speech content or whether it is true or false, right or wrong, and all speech is considered equal "ideas" and "opinions."[47]

However, in recent years, with events such as Russian intervention in U.S. elections, the Brexit referendum, the COVID-19 pandemic, the proliferation of hate speech, and the Russia-Ukraine conflict, there has been an increasing demand for platforms to regulate the content of speech.[48] The tension between the need for regulation and the prohibition of regulating speech content has become more prominent. This tension is mainly manifested in two aspects. On one hand, in the controversies surrounding fake news, online rumors, and cognitive wars, the most prominent issue is the authenticity of speech content. The original assumption of classic freedom of speech was to let true and false, good and bad speech compete freely in the marketplace of ideas; regulation would disrupt that competition. However, people now realize that allowing misinformation and disinformation to spread unchecked will seriously undermine competition in the marketplace of ideas and ultimately lead to "bad money driving out good money." On the other hand, in controversies such as hate speech, the most prominent issue is discriminatory speech against minorities or vulnerable groups, which involves tension between freedom of speech and equality (especially racial and gender equality).[49] Classic freedom of speech advocates for allowing different voices to compete freely on such issues rather than banning hate speech.[50] However, the increasing severity of hate speech in recent years has also made the public aware that allowing minorities and vulnerable groups to "compete freely" with the majority and dominant groups is akin to allowing elementary students to compete freely with professional boxers, which will continue to harm minorities and vulnerable groups and may eventually lead them to withdraw from public discourse.[51] Therefore, restrictions must be imposed on such content.

Second, reconstructing a hierarchical and classification system of freedom of speech. Breaking the prohibition of "content-based" regulation does not mean subjecting all content to "content-based" scrutiny, but rather strengthening regulation on content that truly concerns public discourse and public interests. This requires breaking the traditional hierarchical and classification system of freedom of speech. Classic freedom of speech is a two-tier hierarchical system: Political speech enjoys the highest protection in the first tier, while other types of speech are in the second tier with relatively weaker protection.[52] For the first tier of political speech, there should be no regulation or intervention in principle; for the second tier of other speech, different degrees of regulation may be applied.[53] The classification of classic freedom of speech refers to the categorization of certain speech as low-value due to its distance from political speech and public discourse, making it subject to content regulation.[54] The list of low-value speech is strictly

[47]*See generally* Robert Post, Democracy, Expertise, and Academic Freedom: A First Amendment Jurisprudence for the Modern State (2013).

[48]*See generally* Jack Goldsmith & Andrew Keane Woods, *Internet Speech Will Never Go Back to Normal*, The Atlantic (Apr. 25 2020), https://www.theatlantic.com/ideas/archive/2020/04/what-covid-revealed-about-internet/610549/.

[49]*See generally* Jeremy Waldron, The Harm in Hate Speech 1–17 (2014).

[50]*See generally* C. Edwin Baker, *Harm, Liberty, and Free Speech*, 70 S. Cal. L. Rev. 979 (1997). Abrams v. United States, 250 U.S. 616 (1919) (Holmes, J., dissenting).

[51]Waldron, *supra* note 49, at 16.

[52]Cass Sunstein, Democracy and Problems of Free Speech 121–165 (1993).

[53]*Id.* at 121–65.

[54]*See* Genevieve Lakier, *The Invention of Low-Value Speech*, 128 Harv. L. Rev. 2166 (2015).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

controlled, with only a few categories included, currently including: fighting words, obscenity, child pornography, libel, profanity, and commercial speech.[55]

The conflict between content moderation and the traditional hierarchical and classification system is reflected in the fact that under the traditional two-tier hierarchical and classification system, political speech and public discourse are at the core, and therefore should not be subject to regulation. However, in reality, misinformation, hate speech, and online violence are most rampant and prominent in political speech and public discourse, leading to a stronger demand for regulation. In other words, both the old and new approaches of regulation consider political speech and public discourse the core of freedom of speech, but they prioritize in contrasting ways—what platforms need to regulate the most is precisely what old freedom of speech laws and theories do not allow to be touched. False information and hate speech often erupt during major events, which belong to the first tier of political speech and public discourse, which increases resistance to regulation. Therefore, reconstructing the hierarchical and classification system means that because political speech and public discourse are important, regulation must be applied to ensure their proper functioning. Also, expanding the categories of low-value speech to include new categories such as fake news and hate speech, and gradually developing complementary principles and standards.

D. Challenges and Responses: Online Violence as a Case Study

The previous section discussed the challenges posed by new technologies such as artificial intelligence to content moderation, many of which are universal rather than unique to China. This section will analyze one of the most prominent issues in content moderation in China currently : online violence (网络暴力, *Wang Luo Bao Li*). By using online violence as a case study, this section will outline the problems that Chinese content moderation faces in the context of new technologies. It will also address possible responses.

I. What is Online Violence?

Online violence is a new phenomenon, often referred to as "wang bao" in Chinese. The term "wang bao" has become widely used in the Chinese cyberspace—for example, people often say, "I have been wang bao (online bullied)." Of course, whether they have actually been a victim of online violence is another question.

Yet the frequent use of the term does not necessarily mean that there is a clear definition and standard for online violence. Generally speaking, online violence is considered to be a phenomenon where individuals maliciously attack, insult, defame, harass, or threaten others through social media platforms, causing harm to the dignity, reputation, mental health, and daily life of others. It is characterized by fast dissemination, wide influence, and strong concealment, often causing serious harm to the victims. Online violence may usually take the form of (a) verbal attacks; (b) malicious rumors; (c) human flesh search; or (d) group attacks. Verbal attacks are the most common, and include the use of insulting language, profanity, and derogatory comments towards others on social media, comment sections, forums, et cetera, or making remarks that are aggressive or discriminatory, belittling others' character, appearance, abilities, et cetera. Malicious rumors constitute intentionally fabricating false information to discredit someone's reputation, which may involve personal privacy, moral qualities, professional abilities, et cetera, causing negative impacts on the individual's social image and personal life. Human flesh search means obtaining someone's real identity information, such as name, address, workplace, phone number, et cetera, without their consent through various online means and publicly disclosing this information on the internet, causing significant distress to the individual's life and potentially

[55]Chaplinsky v. New Hampshire, 315 U.S. 568, 572 (1942); United States v. Stevens, 559 U.S. 460, 481 (2010).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

endangering their personal safety. Group attacks occur when people target specific individuals or groups for concentrated attacks and harassment, creating a strong public opinion pressure on the victims through continuously posting negative information, malicious comments, et cetera—essentially forming a "gang attack" on the victim. In addition, some may also include hate speech, malicious reports (恶意举报, *E Yi Ju Bao*), et cetera. However, as will be analyzed later, these definitions may have a tendency to be expanded or even "stigmatized."

In 2022, the Office of the Central Cyberspace Affairs Commission issued the "Notice on Strengthening the Governance of Online Violence" (referred to as the "Notice" hereafter), marking the first appearance of "online violence" in public official documents. The Notice defines online violence as "the concentrated dissemination of insults, libel, rumors, violation of privacy and other unfriendly information against individuals which infringe upon the legitimate rights and interests of others, disturb normal internet order."[56] In 2023, the Supreme People's Court, the Supreme People's Procuratorate, and the Ministry of Public Security issued the "Guiding Opinions on Punishing Online Violence and Criminal Acts in Accordance with the Law" (referred to as the "Guiding Opinions" hereafter).[57] Although the "Guiding Opinions" did not provide a formal definition of "online violence," it described it as

> [a]cts of online violence involving the unrestrained dissemination of insults, defamation, defamation, privacy violations and other information against individuals on the Internet, degrading the character of others, damaging reputations and causing serious consequences such as "social death," even mental disorders, suicide and other severe outcomes; disrupting internet order, destroying the internet ecology, creating a pervasive atmosphere of hostility in cyberspace, seriously affecting public sense of safety.[58]

Currently, the most "official" definition of online violence comes from the "Regulations on the Governance of Online Violence Information" issued in 2024 (hereinafter referred to as the "Regulations").[59] Article 32 of the Regulations defines online violence as "illegal and harmful information that is concentratedly disseminated through the internet in the form of text, images, audio, videos, etc., containing insults, libel, incitement of hatred, coercion, violation of privacy, as well as accusations, mockery, disparagement, discrimination affecting physical and mental health."[60]

According to the legal hierarchy in China, the Regulations are ministry rules. Above departmental regulations, there are laws (statutes) passed by the National People's Congress and its Standing Committee, as well as administrative regulations passed by the State Council. However, there are currently no specific regulations on internet violence in statutes and administrative regulations. The Criminal Law has provisions for crimes of insult and defamation, the Chinese Civil Code protects the right to reputation and privacy, and the Public Security Management Punishment Law has provisions for public insult or defamation based on false information, but none of these specifically targets online violence. Therefore, some voices in the

[56]Guānyú qièshí jiāqiáng wǎngluò bàolì zhìlǐ de tōngzhī (关于切实加强网络暴力治理的通知) [Notice on Strengthening the Governance of Online Violence] (promulgated by Off. Cent. Cyberspace Affs. Comm'n, Nov. 4, 2022), https://www.cac.gov.cn/2022-11/04/c_1669204414682178.htm.

[57]Guānyú yīfǎ chéngzhì wǎngluò bàolì wéifǎ fànzuì de zhǐdǎo yìjiàn (关于依法惩治网络暴力违法犯罪的指导意见) [Guiding Opinions on Punishing Online Violence and Criminal Acts in Accordance with the Law], Sup. People's Ct. Judicial Interpretation (Sup. People's Ct. Sep. 20, 2023), https://www.spp.gov.cn/spp/xwfbh/wsfbt/202309/t20230925_629255.shtml#2.

[58]*Id.*

[59]Wǎngluò bàolì xìnxī zhìlǐ guīdìng (网络暴力信息治理规定) [Regulations on the Governance of Online Violence Information] (promulgated by Off. Cent. Cyberspace Affs. Comm'n, June. 14, 2024, effective Aug. 1, 2024), https://www.cac.gov.cn/2024-06/14/c_1720043894161555.htm.

[60]*Id.*

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

Chinese academic community have advocated for separate legislation specifically addressing online violence.[61]

From the perspective of content moderation, as mentioned earlier, traditional Chinese content moderation has typically forbidden certain categories of content, often referred to as "X no-nos." The "Administrative Measures for Internet Information Services" specifies nine categories of prohibited content, including:

> 1. Opposing the basic principles established by the Constitution; 2. Endangering national security, leaking state secrets, subverting state power, and undermining national unity; 3. Harming national honor and interests; 4. Inciting ethnic hatred and discrimination, and undermining national unity; 5. Undermining state religious policies, promoting cults and feudal superstition; 6. Spreading rumors, disrupting social order, and destabilizing society; 7. Spreading obscenity, pornography, gambling, violence, murder, terrorism, or inciting crime; 8. Insulting or defaming others, infringing on their legitimate rights and interests; 9. Containing other prohibited content under laws and regulations.[62]

Online violence may overlap with the fourth, sixth, and eighth categories, but it is not completely consistent. This incongruity reflects the fact that online violence is still a relatively new phenomenon which has not yet been fully incorporated into the existing content moderation system by higher-ranking laws and regulations.

II. Challenges and Responses

The main challenges that online violence poses for content moderation today are as follows.

First, online violence is difficult to define clearly. Online violence is not physical violence. Represented by the Chinese Criminal Law, there are numerous provisions targeting physical violence. For example, article 236 of the Criminal Law stipulates that those who rape women using violence, coercion, or other means shall be sentenced to three to ten years in prison. Likewise, article 263 stipulates that those who rob public or private property using violence, coercion, or other methods shall be sentenced to three to ten years in prison and fined; and those who commit home robbery, robbery on public transportation, or use particularly cruel methods shall be sentenced to ten years, life imprisonment, or death, in addition to fines or confiscation of property. Furthermore, there is the crime of intentional injury, as stated in article 234, which imposes three years of imprisonment, detention, or control for intentionally harming someone's body, with higher penalties for severe injuries or death caused by extreme cruelty. In real-world cases, there will always be disputes about whether a behavior conforms to the standards established by the Criminal Law or the Public Security Management Punishment Law. For the public, however, deliberate harm to others, physical assault, robbery, and rape are generally considered "violence" or "violent behavior" without much controversy.

Such is not the case for "online violence." For many, online violence does not fall into this familiar and traditional sense of "violence." For example, in the "Guiding Opinions on Punishing Criminal Acts of Online Violence in accordance with the Law," the Supreme People's Court and the Supreme People's Procuratorate often punish online violence as crimes such as defamation, insult, and infringement of personal information.[63] While defamation, insult, and infringement of personal information are crimes, they do not seem to fit the common understanding of "violence."

[61] *See generally* Liu Yanhong, Research on the Legal Governance of Online Violence (网络暴力治理法治化研究) (Law Press China 2023).

[62] Hùliánwǎng xìnxī fúwù guǎnlǐ bànfǎ (互联网信息服务管理办法) Measures for the Administration of Internet Information Services (promulgated by Off. Cent.Cyberspace Affs. Comm'n, 2024).

[63] *See generally supra* note 57.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

To a great extent, the concept of "online violence" is not only vague but also misleading in many ways. It gives the impression that it simply transfers real-world violence to the online space, but actually they are entirely different.

More importantly, as analyzed earlier, in the era of artificial intelligence, traditional system of freedom of speech and classical doctrines face huge challenges, and online violence is one such challenge. It is a completely new phenomenon and category. The "Guiding Opinions" provide hints by treating online violence as defamation, insult, and infringement of personal information crimes.[64] Elements involving defamation in online violence make it somewhat similar to fake news. Elements involving insult in online violence make it somewhat similar to hate speech. As discussed earlier, fake news and hate speech separately pose significant challenges to content moderation today. Online violence combines the common characteristics of fake news and hate speech, making it even more complex.

However, it is not advisable to dismiss "online violence" outright. There exists a "double standard" in dealing with online violence: when expressing one's own views, individuals often advocate for exercising freedom of speech, but when criticized by others, they claim to be a victim of "online violence." Online violence is actually an upgraded version of human flesh search. As argued in another Article, human flesh search certainly includes illegal elements, but it also includes many positive aspects.[65] Especially, human flesh search as well as "online violence" may involve ordinary citizens expressing their opinions on issues they care about. In a sense, it is a form of online free speech mass movement pioneered and embraced by millions of Chinese citizens. The key is to make a fine distinction and differentiate between the few lawbreakers and the majority of ordinary people just expressing their views. There is a Chinese saying that goes, "[w]hen pouring out dirty bathwater, be careful not to throw out the baby as well." When dealing with online violence, the same attitude should be adopted.

Second, online violence presents new challenges to platforms. Undoubtedly, platforms must still bear the "principal responsibility" for content moderation when dealing with online violence. Faced with a large volume of content, platforms must also utilize various technologies for content moderation. In the "government-platform-citizen" triangle, only platforms possess such technological capabilities.

Article 7 of the "Regulations on the Governance of Online Violence Information" stipulates: "Network information service providers should fulfill the principal responsibility for managing the content of online violence information, establish a sound governance mechanism for online violence information, and complete systems for user registration, account management, personal information protection, information release review, monitoring and warning, identification, and disposition."[66] This includes various aspects such as establishing mechanisms, user registration, account management, information release review, monitoring and warning, identification, and disposition—almost all of which require the use or reliance on new technological methods. Legislators also demonstrate a positive embrace of new technologies in content moderation. For example, Article 12 specifically mentions the use of new technologies such as artificial intelligence in alerting and detecting online violence incidents:

> Network information service providers should refine the classification standards and rules for online violence information under the guidance of the National Internet Information Office and relevant departments of the State Council, establish and improve a feature database and sample library of online violence information, and strengthen the identification

[64]*See generally supra* note 57.

[65]*See generally* Yilu Zuo, *A Single Spark Can Start a Prairie Fire: China's Human Flesh Search and Freedom of Expression*, 11 Vienna J. Int'l Const. L. 43 (2017).

[66]*See supra* note 59, at art 7.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

> and monitoring of online violence information through a combination of artificial intelligence, big data, and manual review.[67]

Similarly, Article 23 requires platforms to establish protective mechanisms for potential victims:

> Network information service providers should establish sound functions for protecting against online violence information, providing convenient options for users to block unfamiliar users or specific users, control the visibility scope of their own information, prohibit reposting or commenting on their own information, and improve private message rules, offering convenient options for users to only receive private messages from friends or refuse all private messages, encouraging the provision of smart blocking of private messages or customization of private message block terms.[68]

Many platforms, including Douyin, have established a "one-click anti-online violence" feature.[69] By clicking a button, users can access three main functions: (1) within 7 days, they will not receive private messages, danmaku (弹幕, *Dan Mu*), or comments from users outside their selected group; (2) by enabling "fan list blocking," their account will not be displayed in others' follow and fan lists, preventing online abusers from accessing user information through follower lists and further protecting user privacy; (3) users can use the "batch reporting of private message online abuse" feature to submit private messages, comments, and other content received since the online abuse with one click, providing a convenient way to quickly report online abuse situations to the platform for immediate action.[70]

However, it is worth noting that under the government-enterprise relationship in China, Chinese platform companies will undoubtedly be more "cooperative" with the government than in the West.[71] Still, it would be too simplistic to assume that the government is the only force pulling the strings. On the one hand, Chinese platforms, like Western platforms, will engage in content moderation based on their own interests and ideologies; on the other hand, Chinese platforms may "escalate" their measures. This means that concerning government laws and policies, platforms are likely to engage in excessive moderation, whether for liability considerations or other reasons. The issue of excessive platform intervention in online violence and other matters is equally worth paying attention to.

This is when the principle of "no-dispute" comes into play. In cases of online violence and social hot-button issues, there are often conflicts between different groups and values. Whether it is the government or platforms, they should adhere to a "no-dispute" stance, neither supporting nor opposing any specific position, but instead adopting a consequentialist and pragmatic approach, only managing content that may lead to serious consequences. In the "Guiding Opinions" issued by the Supreme People's Court and the Supreme People's Procuratorate, a more consequentialist understanding of online violence was adopted. It stated that the constitutive elements of online violence include "causing serious consequences such as "social death" of others, mental disorders, or even suicide."[72] However, the subsequent "Provisions on the Governance of Online Violence Information" removed this requirement for consequences. To a certain extent, a more consequentialist approach could be taken in defining and regulating online violence.

[67] *See supra* note 59, at art. 12.

[68] *See supra* note 59, at art. 23.

[69] *Douyin Shequ Zilv Hongyue* (抖音社区自律公约) [*Douyin Community Self-Discipline Convention*], https://www.douyin.com/rule/policy#heading-0.

[70] *Douyin Guanyu Daji Wangluo Baoli Yanlun de Guifan* (抖音关于打击网络暴力言论的规) [*Douyin Rules on Combating Online Violence*], https://www.toutiao.com/article/7260311332904354304/?wid=1772501771814.

[71] *See* Lizhi Liu, From Click to Boom: The Political Economy of E-Commerce in China (2024) (discussing the government relationship more deeply).

[72] *See supra* note 57.

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

Third, the medium-based model faces the possibility of being completely abandoned. This challenge mainly arises from two aspects: the trend of media convergence and the emergence of "personalized law."[73] Respecting media convergence, the existence and rationality of medium-based content moderation are based on the premise of isolation between different media. Traditionally, content from live events, newspapers, broadcasting, movies, and the internet belonged mainly to their respective mediums and did not significantly or automatically flow to other mediums. For example, TV programs would not be broadcast in cinemas, and the movies currently playing in cinemas would not be broadcast on television simultaneously or immediately. However, the internet has evolved from being one of many parallel mediums to becoming the foundation for all media and content. In other words, because all content now ultimately reaches platforms, should different content-moderation standards still be applied based on different mediums, or should there be a uniform standard? If a uniform standard is the preference, what should this new standard be?[74]

Notably, big data, algorithms, and artificial intelligence bring about the possibility of "personalized law." The foundation of the medium-based model actually lies in the most traditional legal thinking—categorization and contextualization.[75] Different media represent different categories and contexts. Discussions in legal theory about "rules" and "standards" are manifestations of categorization and contextualization.[76] In situations where regulators do not have complete information and sophisticated tools, regulating through categorized rules and standards based on different categorization and contextualization becomes the best choice. Specifically, inadequate predictability and interactiveness are two major factors constraining traditional regulations. Predictability refers to the limitation of the law in generalizing predictions about the behavior and reactions of the majority of people through one-size-fits-all rules and standards, while interactiveness refers to the inability for regulators and the regulated party to establish real-time connections and instead rely on legal means as a one-time form of communication.[77] For example, in road traffic safety laws, regulators cannot monitor or control every vehicle and driver, so regulation relies on rules like speed limits and standards like "drive safely." The setting of different speed limits for highways, national roads, and provincial roads is based on different categorization and contextualization for regulation. Content moderation based on the different technologies and social characteristics of various media follows the same logic. Live events, printed media, broadcasting, movies, and the internet represent conceptualized different categorization and contextualization. Under each category and context, content moderation develops more refined rules and standards.

New technologies such as big data, artificial intelligence, and algorithms enable the possibility of creating profiles and analysis for each audience member. In the future trend of "personalized law," if "personalized content moderation" indeed emerges, it means that content moderation can directly affect each audience member. In other words, content moderation no longer needs to assume that a father listening to the radio or watching TV might be accompanied by his child; rather, it can utilize data collection, analysis, profiling, and real-time interactions to understand, for instance, whether a child is sitting beside the father, the father's attitude towards minors accessing certain content, et cetera, and then precisely deliver or block content. Furthermore, the applications of "personalized content moderation" extend beyond protecting minors. For adult audiences, preferences such as what type of content they want or do not want to see, which public figures they want or do not want to encounter, and specific voices or viewpoints they desire or do not desire can be customized. If such a day indeed arrives, distinguishing between mediums may

[73]Omri Ben-Shahar & Ariel Porat, Personalized Law: Different Rules for Different People 20–23 (2021).
[74]*See generally* Zuo, *supra* note 11.
[75]*See supra* note 73, at 20–23.
[76]*See generally* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 Duke L.J. 557 (1992).
[77]*See generally* Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards*, 92 Ind. L.J. 1401, 1403 (2017).

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press

no longer be necessary except for live events, as differentiation has precisely narrowed down to each individual. Therefore, to some extent, the advent of personalized law may signal the demise of the "medium-based" model.

If the challenge brought by media convergence is "how to handle it," the introduction of "personalized law" primarily raises a fundamental and normative question: is this "brave new world" something people aspire to? In the eyes of advocators of "personalized law," more precise and customized regulation does indeed surpass traditional regulations. However, concerning content moderation, criticism may arise from two aspects: first, "personalized law" necessitates the collection and analysis of vast amounts of data, potentially infringing on privacy and personal information; second, excessive personalization and customization may lead individuals to only access information they want to see, making people increasingly narrow-minded and extremist, resulting in phenomena like information bubbles, echo chambers, and group polarization.[78]

E. Conclusion

As emphasized earlier, this Article discusses the challenges brought by the age of artificial intelligence—not just AI technology itself—to content moderation in China. This means that, in terms of time, the issues facing China's content moderation are an accumulation of those from previous eras and various technologies. In other words, China's content moderation today may need to address all the problems that Western countries have experienced in the 19th, 20th, and 21st centuries simultaneously. In terms of space, the challenges facing China's content moderation include both universal challenges and those unique to China. Therefore, this Article provides an outline of content moderation in the AI age in China, hoping to offer a framework for more comprehensive discussions in the future.

Acknowledgements. The author declares none.

Competing Interests. The author declares none.

Funding Statement. No specific funding has been declared in relation to the Article.

[78] *See generally* Cass Sunstein, Republic.com 2.0 (2007).

Cite this article: Zuo Y (2025). China's Content Moderation in the Age of Artificial Intelligence. *German Law Journal* **26**, 1300–1319. https://doi.org/10.1017/glj.2026.10195

https://doi.org/10.1017/glj.2026.10195 Published online by Cambridge University Press