Shifting platform governance: examining participatory content moderation on a Chinese platform Bilibili

Shang, Z.

INFORMATION, COMMUNICATION & SOCIETY
https://doi.org/10.1080/1369118X.2025.2520004

Routledge
Taylor & Francis Group

Shifting platform governance: examining participatory content moderation on a Chinese platform Bilibili

Zheyu Shang

Institute for Area Studies, Leiden University, Leiden, the Netherlands

ABSTRACT

In China, digital platforms are held accountable for monitoring online content under current laws and regulations. Legal pressure and the sheer amount of user-generated content require platforms to improve their moderating ability. Bilibili, a prominent Chinese video-sharing platform, established a Disciplinary Committee to mobilize users to review comment violations and foster a positive community ambiance. This study examines Bilibili's content moderation mechanism defined as participatory content moderation. Using app walkthrough method, this article explores the features and potential risks of this new mechanism. The findings suggest that participatory content moderation provides a more accessible, transparent, and collaborative model, shifting the role of users in platform governance. However, the system's reliance on major user preferences and the influence of national regulations introduce complexities that affect its inclusivity and democratic potential.

ARTICLE HISTORY
Received 30 May 2024
Accepted 10 June 2025

KEYWORDS
Content moderation; platform governance; Bilibili; participatory culture; danmu

1. Introduction

While users immersing in the bright side of cyberspace with free expression and interaction, a dark facet exists where unfriendly and harmful content like disinformation, child pornography, hate speech, and violent content poses threats to both the integrity of online communities and users' well-being.

Content moderation, although not always visible to ordinary users, remains an essential part of platforms. Essentially, platforms as data infrastructures are designed to digitize and monetize users' published content and online interactions (Poell et al., 2019; van Dijck et al., 2018). From a business standpoint, they have to moderate content to retain users and uphold a positive image to advertisers and the public (Gillespie, 2018).

From the perspective of legal requirements, unlike US-based social media platforms that are sheltered under safe harbour provisions, Chinese platforms are held accountable for managing content shown on their platforms. Legislative mandates like the Regulation on Internet Information Service of the People's Republic of China (2004), the

CONTACT Zheyu Shang zheyu.shang@outlook.com Institute for Area Studies, Leiden University, Herta Mohr, Witte Singel 27A, Leiden 2311 BG, the Netherlands

Cybersecurity Law of the People's Republic of China (2016), and Provisions on Ecological Governance of Network Information Content (2019) clarify the responsibility of online platforms to strengthen content governance, detect and remove illicit content, and promote a positive cyberculture. The Cyberspace Administration of China further tightened the regulations on user comments in 2017 by introducing the Provisions on the Management of Internet Post Comments Services, which necessitated a more careful examination of user comments. According to the provisions, platforms as internet service providers are required to monitor users' comments, improve comment management, build a professional team for content review, and cooperate with governments' supervision. Platforms that violate or are failed to meet these requirements will be summoned or punished by cyberspace administration authorities. These detailed regulations on online content review spurs platforms to refine content moderation strategies. Consequently, Chinese platforms face an obligation to rigorously moderate content, guarding against legal repercussions and potential shutdowns.

Situated within this intricate regulatory landscape is Bilibili, one of the most popular online video-sharing platforms in China with 314.5 million active users (Bilibili, 2023). It is distinguished by its danmu function (弹幕, also known as bullet curtain or danmaku), which allows users to send comments overlaid on the video footage (Figure 1). This function provides a new conduit for users' creative expressions, creates a co-viewing experience by showing asynchronous danmu comments in a synchronic manner, and adds extra entertaining values to videos (Chen et al., 2017; He, 2022; Schneider, 2021). By the end of 2021, the total amount of danmu comments has reached over 10 billion and is still increasing by nearly 2 billion every year (Bilibili, 2021). The sheer amount

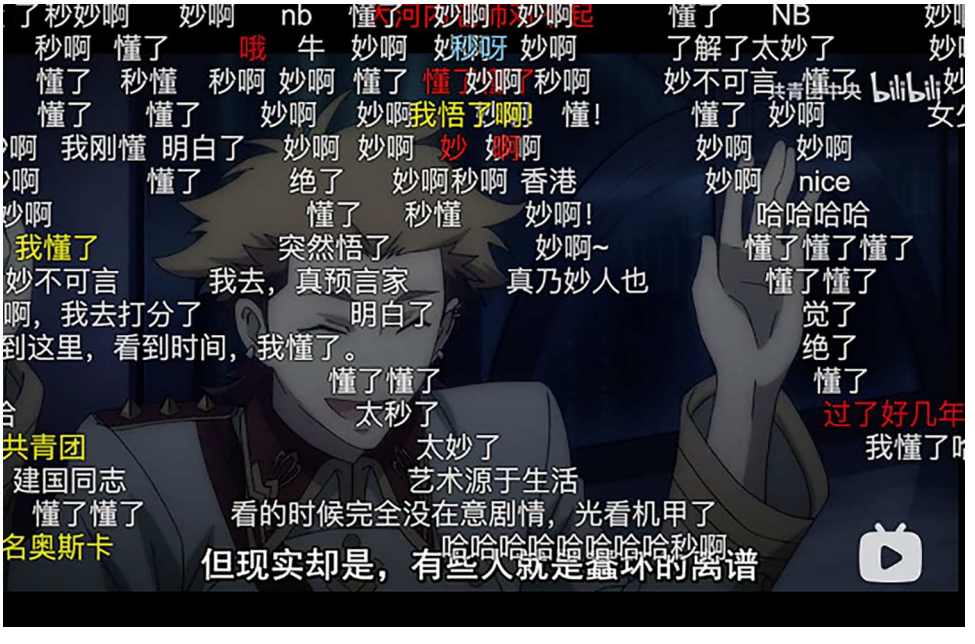

Figure 1. Danmu comments on Bilibili. https://space.bilibili.com/20165629/video [Accessed 07 May 2021].

of danmu comments creates an active online environment for users but also places burdens on Bilibili's content moderating ability. As a platform that hosts user-generated content, Bilibili is legally required to moderate video content and danmu comments. However, little attention has been devoted to how danmu comments are reviewed, filtered, and removed behind the screen.

In response to the increasing user-generated content and regulatory pressure, Bilibili has taken several measures such as developing algorithms and building a monitoring team of 3800 employees for manual review (Bilibili, 2023). Yet, challenges persist. First is the accuracy of automatic detection. Rooted in the ACG culture (Animation, Comics, and Games), Bilibili has become a cluster of subcultures and young internet users. The nuanced linguistic dynamics cause difficulties for automatic detection. The manual check with employees who are familiar with these cultural tropes may help to solve this problem, but it leads to the second problem: the problem of scale. The huge amount of user-sent danmu comments increases the cost and difficulty of manual review.

To address the challenges of scale and inaccurate automatic detection, most platforms around the world have adopted a flagging system that enables users to report inappropriate content. Bilibili, however, has taken an additional step. Since 2017 when the Provisions on the Management of Internet Post Comments Services was carried out, Bilibili established a Disciplinary Committee (风纪委员会, DC) to engage its users in assessing reported content violations and general danmu atmosphere through a jury-verdict system. So far, this innovative approach has been widely embraced by many Chinese platforms, including but not limited to Weibo, Douyin, Zhihu, and Meituan.[1]

While this mechanism is prevalent among Chinese platforms, the underlying platform design and the evolving dynamics between users and platforms during the moderation process remain underexplored. Examining the case of the disciplinary committee operating system offers insights into users' role in content moderation and the strategies employed by platforms to mobilize users for voluntary moderation work.

This study aims to investigate the case of Bilibili to bring the innovative content moderation defined as participatory content moderation mechanism into research scope and fill the research gap in users' roles in platform governance and in the Chinese context. In this article, I will delve into the design of the disciplinary committee and the overarching architecture for participatory content moderation, guided by the subsequent research questions:

- What is participatory content moderation? How does Bilibili leverage participatory culture to involve users in content moderation?
- What are the features and potential risks of the participatory content moderation mechanism?
- In what ways does Bilibili's content moderation strategy alter the platform-user dynamic in platform governance?

2. Literature review

2.1. Content moderation

Content moderation is integral to platform operation. Gillespie (2018) delineated it as essential for platforms to protect users and establish a positive image to the public and

advertisers. While platforms seldom produce content directly, they play a role in its production and distribution through content moderation. This process not only involves removal of inappropriate content but also influences content visibility, user consumption patterns, and the discipline of content producers (Gillespie, 2018; Zeng et al., 2022). As platforms increasingly mediate cultural production, the broader implications of content moderation on the cultural industries become increasingly apparent (Nieborg & Poell, 2018; Poell et al., 2021).

Based on studies of Western-based platforms and websites like Facebook, Reddit, Twitch, Twitter, and Wikipedia, scholars have indicated three common mechanisms:

Commercial content moderation (CCM) is commonly utilized and refers to the platform-driven, paid process of assessing user-generated content by human reviewers (Roberts, 2019). Given that social media platforms are saturated with content rooted in human values and culture, CCM leverages human labour to effectively manage reported materials and identify inappropriate content. In practice, this mechanism is obscured by platforms' opaque moderation practices, leading to ambiguity, misunderstanding, and distrust among users (Cook et al., 2021; Gillespie, 2018; Suzor et al., 2019).

Community moderation, embraced by platforms such as Reddit and Twitch, delegates part of administrative duties and power to certain users within online groups or communities, giving them more autonomy in making internal rules and regulating others and content in online communities (Cook et al., 2021; Matias, 2019; Seering, 2020; Thach et al., 2022). While this approach is more transparent and prioritizes user autonomy, tensions can arise between user moderators and platforms, as seen in instances like the 'Reddit blackout' (Matias, 2016). Moreover, the moderation teams formed by users tend to be fixed and limited within a small scale.

Algorithmic content moderation (ACM) is designed to process vast amounts of content quickly. ACM uses matching and predictive systems to detect potential content violations by either comparing them with existing violation dataset or training artificial intelligence (AI) to review content (Gorwa et al., 2020). Despite its efficiency, ACM has shortcomings, particularly in reviewing content in complex language and context that requires nuanced understanding (Caplan, 2018; Gillespie, 2018, 2020; Gillespie et al., 2020). Similar to CCM, ACM is also directly organized by platforms and suffers from transparency issues (Gorwa et al., 2020).

Although content moderation practices evolve into different mechanisms to meet various needs, existing mechanisms are not infallible in terms of scale, accuracy, and transparency. The border between platform-led and user-led moderation is evident, while collaboration remains minimal. This current dynamic leaves the user's position ambiguous, making their contribution in the broader moderation landscape unclear and underexplored.

Some scholars have recommended decentralizing responsibility and fostering collaboration between platforms, public institutions, and users (Caplan, 2023; Helberger et al., 2018). However, the practical blueprint to achieve such cooperative responsibility and utilize user intelligence remains ambiguous. While platforms' algorithms can underperform in recognizing intolerant speech lacking toxic language or pro-democratic speech containing incivil expressions (Oh & Downey, 2024), users tend to be adept at distinguishing between incivility and intolerance, but they show low support for content moderation actions, such as content removal or user bans (Pradel et al., 2024).

Additionally, user groups are diverse with their own linguistic norms and cultural references, which might trigger automatic detection and result in erroneous content removal (Dias Oliva et al., 2020; Haimson et al., 2021). This highlights the importance of understanding user perceptions and language use within its social and cultural context. Notably, even though platforms like Twitter and Facebook are reported to initiate the Community Notes and the Community Council to include users in content moderation tasks as a supplement to their ACM and CCM (Eldon, 2010; Twitter, n.d.), they appeared to make little progress so far.

While much existing research in platform governance focuses on major Western platforms, the platform-designed and user-engaged content moderation mechanism employed by Chinese platforms are often overlooked. Some Chinese scholars have demonstrated the positive impact of this collaboration on user loyalty and public security and identified the DC on Bilibili as a mediator between the party-state's regulatory pressures and the young-generation users' demands (Chen & Yang, 2023; Mao & Liao, 2020; Zhao et al., 2019). Although they approach the subject from various sociopolitical and management perspectives, little attention has been paid to users' changing role in platform governance and the moderation mechanism itself, including system designs that facilitate such collaboration and the pitfalls of this mechanism.

This research examines this content moderation mechanism with the case of Bilibili. By focusing on this understudied dimension, it provides new insights into a potential content moderation model that engages users with participatory culture and underscores collaboration between platforms and users, which has implications for fostering a balanced digital ecosystem where both users and platforms actively negotiate content moderation standards and shape the content landscape. This study bridges the gap in the existing literature and enhances the comprehension of user-platform interactions in digital spaces.

2.2. Participatory culture and Bilibili

Participatory culture, characterized by active participation, expression, creation, and civic engagement, thrives in both fandoms and digital media platforms (Jenkins, 2009). The concept of participatory culture marks the evolving roles of ordinary internet users. The low barrier for obtaining and sharing information has enabled users to create content online and resulted in a bottom-up participatory culture (Jenkins, 2009; Jenkins et al., 2013; Rheingold, 2008). Previous studies have explored online participatory culture across various sectors, such as platform economy, fandom activities, and civic participation (Rheingold, 2008; Yin & Fung, 2017; Zhang & Mao, 2013). Here, users are no longer deemed as passive audiences, but active prosumers engaging in content creation, online communication, and interactions (Jenkins, 2009; Manovich, 2001; van Dijck, 2009).

In China, Bilibili exemplifies a platform thriving on participatory culture and users prosumption (Chen, 2018, 2020; Han, 2016;Yin & Fung, 2017). The core function of this platform, the danmu function, boosts users' interactions and content (re)production practices from three aspects. First, danmu enriches the viewing experience by creating a co-viewing atmosphere and bringing (re)production dimension into viewing (He, 2022). Although danmu comments are sent asynchronously, they are displayed in a

synchronous manner. This pseudo-synchronous display mechanism enables users to share and respond to information in danmu (Chen et al., 2017; Wu et al., 2019). Second, high visibility and anonymity render danmu an open public space for expression and interaction. In general, danmu as a whole embedded in the video-playing interface has higher visibility than traditional comments in the isolated comment box, though some danmu comments can be hidden according users' personal settings and the system's capacity limit. Meanwhile, the information of danmu senders is not public to other audiences. Anonymity provides an explanation for the greater number of danmu comments compared to traditional comments (Clark, 2012; Miller et al., 2016; Wu et al., 2019). Lastly, the adaptable danmu system affords users' creative expressions. Previous studies have demonstrated the semiotic resources in danmu enable translation practices and the wide use of linguistic memes (Wu et al., 2019; Yang, 2020). The (pseudo-)synchronicity, visibility, anonymity, and creativity of danmu comments bolster users' presumption and meet users' demands for information, entertainment, and social interactions. Additionally, as Bilibili brands itself as a hub for the young generation, the danmu function extends its reach, serving as a cornerstone for community-building by fostering user participation and reinforcing shared sentiments and cultural tropes (Schneider, 2021; Wang, 2022; Wu et al., 2019; Xiang & Chae, 2021).

Yet, participatory culture presents both advantages and challenges for Bilibili. While the danmu function enhances user engagement, it can lead to potentially harmful content and visual clutter, compromising user experience (Chen et al., 2017). The anonymity of danmu can pose challenges for platform governance as anonymity occasionally driving aggressive behaviours (Zimbardo, 1969). In addition, danmu content moderation faces practical problems as enormous danmu comments are highly context-based with complicated and fast-changing internet languages with subcultural elements, which challenge the platform's content moderation ability.

As platform governance and digital justice call for cooperative responsibility, user engagement should not be neglected (Helberger et al., 2018). Participatory culture and user engagement have the potential to bolster a more democratic platform governance for two reasons. Firstly, in essential, participatory culture is closely related to the values of democracy and diversity (Jenkins et al., 2015), which are currently lacking in platform governance. Secondly, participation is different from mere interaction with digital media, instead, it requires individual and collective decisions and actions to build up shared experiences and values (Jenkins et al., 2015). Enabling user participation in platform governance allows users to utilize collective intelligence to contribute to embedding shared values within platform policies and rules.

However, while previous studies have explored participatory culture in content production and circulation, limited studies have investigated its application in content moderation and users as moderators in the content moderation system since this process was traditionally controlled by platforms and hidden behind the screen. To capture the evolvement of participatory culture and online participation, research on users' evolving roles in platform governance and the ways platforms bolster user participation is needed.

Promoting participatory culture in platform governance, especially content moderation, requires efforts from both users and platforms. As Jenkins (2009) highlighted the three aspects of participatory culture: supportive external conditions, users' sense of agency, and knowledge transfer between users, combining participatory culture

with content moderation needs users' active participation as well as platforms' support such as opening the access to moderating process and data, constructing essential operating systems, and providing clear and updated moderation information.

3. Methodology

This research adopts the app walkthrough method to delve into Bilibili's participatory content moderation system designed for user engagement. Defined as a 'step-by-step observation and documentation of an app's screens, features, and flows of activity', the walkthrough method serves as an instrumental means of probing the technological and cultural elements interwoven within an app interface (Light et al., 2018, p. 882). This technique facilitates the detailed examination of human-app interface interactions, technological functions, and the digital ecosystem that guides users' online activities (Duguay & Gold-Apel, 2023). By simulating users' experiences, researchers can explore how platforms direct user behaviours through varied functionalities and design strategies (Chen et al., 2021; Dieter et al., 2019; Kaye et al., 2021).

Since over 90% of Bilibili's active users access the platform through mobile devices (Bilibili, 2022), the mobile app stands as the primary access point for most users. Therefore, the app walkthrough method provides a direct engagement with the interface most users navigate, offering insights into their activities and potential experiences within the system. To capture the intricate nature of the participatory content moderation system, I conducted app walkthrough targeting interfaces of the community centre where users can gain information about content moderation, and the DC, wherein users actively engage in content moderation via a jury-verdict system.

I used my Bilibili account to join the DC. This immersive experience within the DC and its ecosystem provided valuable insights into the platform's moderation dynamics. From 2020 to 2023, I participated in the DC for eight months, during which I moderated 242 cases about user comments, including danmu comments in videos. Most of these videos were entertainment-oriented (106 cases, including content related to games, anime, films, and comedy), while others focused on daily life (92 cases, including videos about pets, food, hobbies, and vlogs). The remaining cases covered global and domestic news (21 cases), knowledge sharing (13 cases), and technology (10 cases).

4. Features of Bilibili's participatory content moderation system

To encourage users' involvement in content moderation, Bilibili has built a content moderation mechanism based on its community and ACG culture. Here, I define it as participatory content moderation which is an accessible and collaborative approach designed and technically supported by the platform, which invites users to engage in detecting and judging content violations based on both their knowledge and adherence to platform rules.

Compared to CCM and ACM where moderation is dominated by platforms themselves, the participatory model promotes transparency and user engagement. Unlike community moderation, which typically operates within smaller online groups and remains relatively independent of the platform in terms of rulemaking and daily operation, participatory content moderation is embedded in platforms' broader moderation

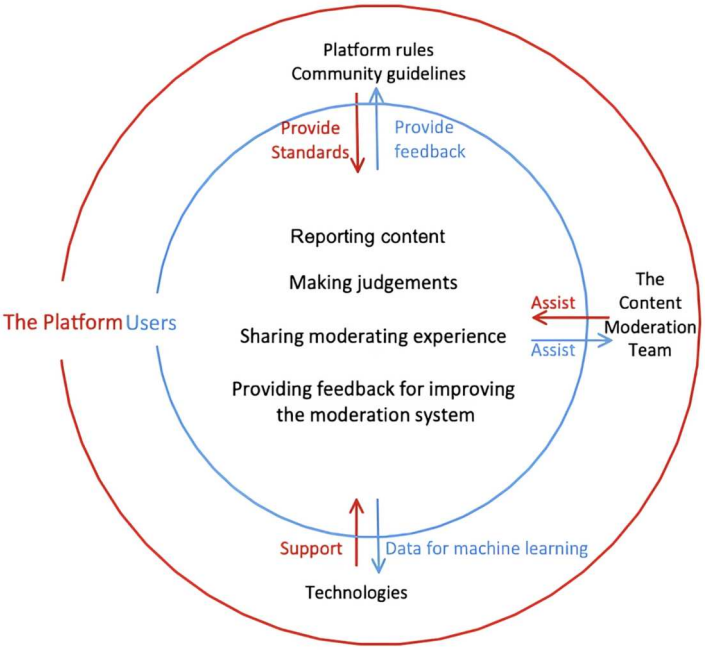

Figure 2. Collaboration between the platform and users in participatory content moderation.

work, guided by platforms' rules, and provides essential data for optimizing platforms' content moderation systems, allowing a large user base to engage more expansively, fostering collaboration and information exchange between users and the platform (Figure 2). On Bilibili, this mechanism operates through the work of the DC and the information release in the community centre.

4.1. Daily practices of the disciplinary committee

As the core at the heart of the participatory content moderation mechanism, the DC is a user-centred moderating organization. Launched in 2017, its primary goal is to uphold a friendly community atmosphere by identifying content violations and making collective decisions on reported content (Figure 3). The name Disciplinary Committee draws inspirations from Chinese campus culture and Japanese ACG culture. In both China and Japan, the role of a disciplinary student leader is prevalent in schools, primarily focusing on maintaining order within the classroom. For Bilibili users who are familiar with either Chinese schools or Japanese ACG culture, this name encapsulates the function and responsibility of this organization and conjures up a more youthful and approachable image rather than serious or bureaucratic images.

According to Bilibili, users can apply to join the DC through the community centre page if they meet specific criteria regarding user active time, frequency of danmu usage, a clean violation history, and the real-name verification. Since recognizing content violations requires familiarity with cyber culture and platform rules, these criteria establish an entrance threshold to select experienced users who are likely capable of handling content violation tasks.

Figure 3. The advertising poster of Bilibili's DC (translated by the author). Source: https://www.bilibili.com/read/cv14023818 [Accessed: 10 Oct. 2022].

While applying to join the committee is simple, maintaining membership demands dedication. Each term lasts one month, and members are expected to remain active for at least 10 days and handle at least 50 cases to qualify for continued participation. DC members generally work on a voluntary basis, though active members can receive immaterial rewards of a specialized virtual badge named 'force for justice' and Bilibili coins.[2]

Members can make collective decisions on potential (danmu) comment violations, through the jury-verdict system. Generally, cases for review originate from two primary sources: the algorithm's random selection from trending videos and user-reported comment violations that exceed the capacity or expertise of the platform's content moderation team.

The interface of the jury-verdict system is divided into two sections as illustrated in Figures 4 and 5: an informational section (including a video module and danmu comments module) and a voting section (featuring a voting module and a discussion module

Figure 4. The interface of the jury-verdict system [Accessed 27 Sep. 2022].

for members to review peer comments). The system is designed to provide contextual information and encourage interaction, with members able to directly play the video to understand the context better, refer to community guidelines, and vote independently or share additional insights to assist others in their decision-making.

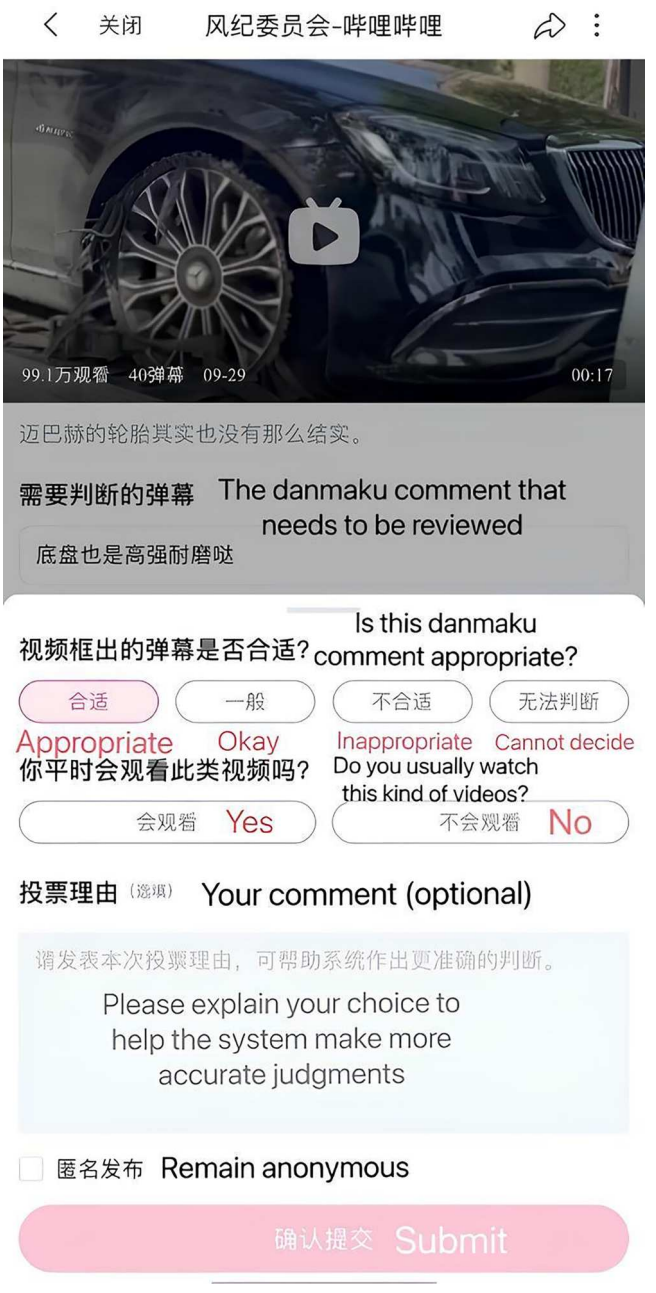

Figure 5. The voting module in the jury-verdict system [Accessed 11th Oct 2022].

A case judgement normally includes around 300 members' votes. Members can monitor the progress of ongoing cases and review the voting outcomes of closed cases in which they have participated. The final decision is determined by a majority vote. However, if members have concerns or disagree with the outcome, they can appeal or report the result to the platform via email. Bilibili occasionally updates the outcomes of controversial cases in its DC weekly reports.

To foster user engagement by cultivating a sense of community, Bilibili leverages both explicit signals (e.g., community centre) and implicit ways like the integration of cultural elements in the interface design and the interactive designs to strengthen interactions between DC members in decision making. This emphasis on community values encourages users to see their moderation efforts as contributions to maintaining the platform's atmosphere, reinforcing the positive perception of content moderation. By involving users in case judgements and highlighting their contribution to the community atmosphere, the platform fosters a sense of user agency in safeguarding the community values, legitimizes the need for content moderation, erasing its negative connotations.

4.2. Four features of the participatory content moderation

Based on the observations from the DC's daily practices, participatory content moderation system involves users' engagement in the entire content moderating chain from detecting possible content violations to making collective decisions on violations. Bilibili's participatory content moderation model can be defined by four key features: shared values and norms, increased transparency and interactivity, lower barriers to participation, and strengthened collaboration between the platform and users.

Shared values and norms

Effective content moderation relies on well-defined standards that distinguish acceptable from unacceptable content. Particularly in participatory systems, users' comprehension and endorsement of these standards are crucial.

From the perspective of platform design, Bilibili uses three strategies to ensure users align with community guidelines. Firstly, unlike many social media platforms that grant immediate access upon registration, Bilibili mandates an entrance exam or an invitation from a high-level user. The entrance exam serves to raise the threshold for joining this community and preserve Bilibili's community culture against fan trolls in previous cases (Zheng, 2019). This helps address the challenge of shared understanding by testing new users on basic platform rules, danmu etiquette, and subcultural norms, which is essential for their further engagement in content moderation. Secondly, the platform provides educational resources, such as videos on community guidelines that have garnered millions of views, further reinforcing user understanding of platform rules. Lastly, Bilibili constantly updates the DC weekly working reports, showcasing selected typical cases from the week, top-performing DC members, and the final results of controversial cases.

These measures help promote a solid foundation for a consistent understanding of content moderation standards and fundamental community values. Additionally, user participation in content moderation further cultivates shared norms through collective practices. This collaboration between users and the platform creates a basis for shared values and practices, making participatory content moderation feasible.

Increased transparency and interactivity

Transparency in content moderation is critical yet often lacking in platform-led systems. Merely disclosing results and the system's content classifications may not guarantee users' understanding and trust in the system. Scholars have argued for interactive

transparency, suggesting increasing it by enhancing user involvement and allowing users to provide feedback on classification standards as a possible solution (Molina & Sundar, 2022).

User engagement and the increased information disclosure from the platform's community centre mentioned above contributes to the increase of transparency in content moderation work. Within the disciplinary committee, members can view others' votes and comments on cases, which not only demystifies the decision-making process but also encourages a richer dialogue about content standards. Moreover, the regular updates in the community centre and the small darkroom that designed for showcasing content violation cases keeps the wider user base informed, which also invites public commentary and supervision (Figure 6)

Compared to traditional platform-led content moderation mechanisms where the accessibility of content moderation for the average users has been limited, the more accessible nature of the participatory content moderation mechanism and the platform support for information publication contribute to a higher level of transparency and interactivity.

Easier participation: low barrier and multiple paths

The most salient characteristic of participatory content moderation is users' participation in content violation judgement on a large scale. Prior to this mechanism, major platforms employ the flagging mechanism, which allows ordinary users to identify and report potential violations, and the flagging system stands as a nexus connecting user

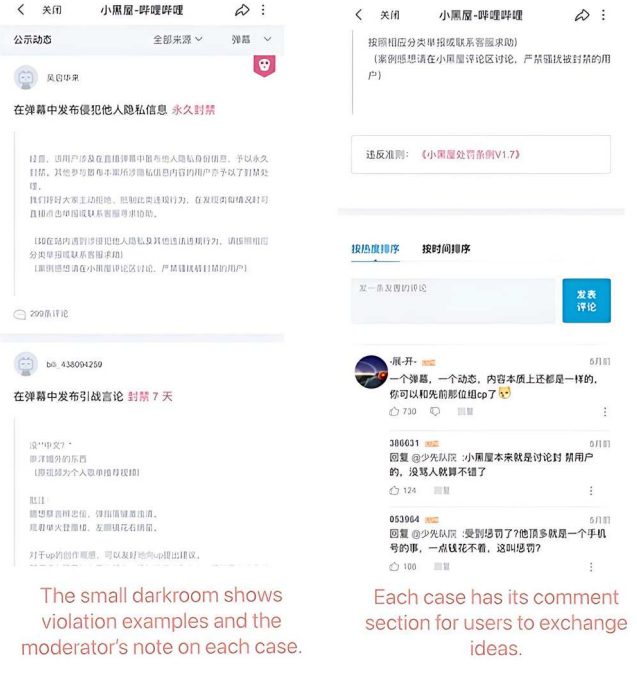

Figure 6. The interface of the small darkroom, showing cases, case results, moderator's notes, and users' comments. Source: https://www.bilibili.com/blackroom/ban [Accessed 16th Aug 2023].

behaviours, platform functionalities, human discernment, algorithmic actions, and wider socio-political and regulatory contexts (Crawford & Gillespie, 2016). Yet, the current flagging mechanism on major platforms has several problems like limiting users' capacity to specify their reports, maintaining opaque decision-making process, lacking feedback on users' reports (Crawford & Gillespie, 2016; Gillespie, 2018). The participatory content moderation system mitigates these issues by reducing technical barriers for users, making the process more accessible. By broadening user involvement from mere reporting to actual decision-making, Bilibili enhances community responsibility and engagement, distinguishing it from platforms where moderation is limited to small, static groups.

Strengthened collaboration between the platform and users

Unlike unilateral traditional models, participatory content moderation mechanism tends to foster a bidirectional exchange where both platform and users contribute to the governance process. This collaborative approach is facilitated by a robust infrastructure that supports active user involvement in refining moderation practices. On Bilibili, the employed moderation team not only responds to user feedback but also adapts to evolving community norms and expectations from DC cases, creating a dynamic ecosystem conducive to sustained user engagement and co-governance.

With these features, Bilibili's participatory content moderation represents a departure from traditional methods, blending both platform-led and user-led mechanisms and leading to a more collaborative ecosystem conducive to active user engagement. As the jury-verdict system interface indicated, cases in the DC provide essential data for refining the platform's AI in content violation detection. It means user engagement becomes an integral layer of the platform's broader moderation approach.

Overall, these features represent a shift from traditional content moderation models to a more collaborative approach that emphasizes user engagement, shared community values, and transparent decision-making. The participatory content moderation has several advantages. From the users' standpoint, it shifts content moderation from an opaque platform-controlled process hidden behind the screen to a more transparent and accessible process while enriching the user experience. For Bilibili, this design paradigm harnesses collective decision-making, engaging a broader user base and mitigating the pitfalls of individual biases.

5. Tensions in Bilibili's participatory content moderation

Participatory content moderation, characterized by users' collective efforts, appears to be a seemingly democratic form of moderation and an expression of civic engagement in online communities. It highlights the value of user participation while downplaying the negative aspects of content moderation. Nevertheless, this system is not without its flaws. In practice, potential risks and tensions still persist.

The first issue arises from user heterogeneity and the platform's inability to match users with their areas of expertise, which may affect the effectiveness of user judgments. While participatory content moderation seeks to leverage users' knowledge to maintain a positive community environment, users have diverse identities and preferences, so it remains uncertain whether users can effectively handle controversial content from various contexts, especially when assigned cases outside their expertise or experience.

According to the author-accessed cases, on average, only 26.17% of DC members are assigned cases that align with their video consumption patterns. This percentage ranges from 17.3% to 41.9% across the 242 cases the author participated in, indicating that most DC members are assigned cases outside their content interests. While diverse voices are needed for a more inclusive community, the platform's inability to balance diversity with specialization may hinder the accuracy and effectiveness of content moderation.

The second issue is the tension between user empowerment and the platform's retention of control over content governance. From the perspective of platform power, previous studies have explored the accumulation and circulation of platform power in content creation and governance, the trend of datafication, and monetization of users' data (Mejias & Couldry, 2019; Poell et al., 2021; van Dijck et al., 2018). Although participatory content moderation promises higher transparency and user engagement, the platform retains control over the formulation of rules, technological infrastructures, and data. Moreover, the datafication of user participation plays a crucial role in this dynamic. Mejias and Couldry (2019, p. 3) argue that the datafication encompasses both 'the transformation of human life into data through processes of quantification, and the generation of different kinds of value from data.' The datafication of users' participation reinforces platform surveillance over users' activities and provides pivotal data for subsequent value extraction at a minimal cost. While the platform claims to engage users, their contributions are processed into data that benefit the platform's power dynamics.

A further tension arises from the limited scope of user engagement in moderating sensitive content. Although Bilibili encourages user involvement in content moderation, the content exposed to users is heavily limited. Political sensitivity and controversial topics like domestic politics, gender, or race are rarely represented in the cases assigned to DC members. Of the 242 cases the author reviewed, 198 cases related to entertainment-oriented content, such as games, anime, films, and comedy, while ideologically and politically sensitive topics were notably absent.

This limitation is likely caused by China's strict regulations on online information, platforms' pre-publication content review where videos containing sensitive or controversial content are often filtered before reaching the platform, and the platform's playful nature that focuses on light-hearted content rather than political discussions. It is crucial to acknowledge that Bilibili users' practices cannot be separated from the broader Chinese context. While participatory content moderation may appear more accessible to ordinary internet users, the visibility and accessibility of content are constrained by national regulations, which tend to be stricter and more sensitive on political and ideological topics than in Western legislative framework. As a result, user engagement is restricted to moderating comments on non-sensitive topics, excluding controversial or politically sensitive content. This restricts the scope of user involvement and perpetuates mainstream ideologies, while potentially side-lining marginalized voices.

Last but not least, the adoption of participatory content moderation does not necessarily guarantee justice or democracy. While the mechanism represents a shift towards collaborative model, it is designed to maintain a balance between platform control and user acceptance rather than striving for justice in content moderation. In Bilibili's case, the platform's system emphasizes user acceptance of content standards over the pursuit of justice. For example, the algorithm captures low-controversy cases randomly

from trending videos, primarily aiming to maintain community cohesion by reflecting the preferences of the majority group of users. The system prioritizes the opinions of the largest user base rather than incorporating more nuanced perspectives on controversial topics. This data-driven approach ensures that content is moderated in line with the broader community's preferences, but it also sidesteps deeper questions of fairness and justice in controversies.

Behind the participatory form, the platform still maintains control over its technological infrastructure and rulemaking. Although most of rules are in line with community values, the platform's community guidelines prioritize the national interest and security over individual freedom and community values to protect the platform from legal risks under China's national security laws. For example, the most prominent rule in the community guidelines is the 'Nine Prohibitions' principle, which emphasizes that content violating the Constitution or other national regulations that threaten national interests and stability is strictly prohibited (Bilibili, n.d.). Rules are made in a top-down manner, and practices of participatory content moderation tend to catalyse users' internalization of platform rules.

By leveraging user participation data, the platform can adjust user roles, modulate the scope of cases, and alter the power dynamics without changing the fundamental system of governance. In this way, participatory content moderation becomes an elastic mechanism, adaptable to the platform's needs while utilizing user knowledge and providing users with a sense of participation and belonging.

As participatory content moderation is in its formative phase, its application on different types of platforms and user perceptions need further exploration. While this research pivots around platform-centric design in the Chinese context, future studies could delve into users' perspectives to understand users' motivations and struggles in this model.

6. Conclusion

The growth of users' creative expression like danmu comments proposes higher demands for platforms to effectively moderate danmu content to ensure a harmonious user environment and circumvent legal pitfalls. This research offers an insight into Bilibili's moderation approach, with a particular emphasis on platform mechanism to engage users in content moderation.

Using the app walkthrough method, this study illuminates the integration of participatory culture and content moderation. The new moderation mechanism actively involves users and foster collective intelligence by offering user-friendly interfaces and delegating the responsibility for reporting and assessing possible content violations. Compared to previous content moderation mechanisms, participatory content moderation is distinguished by its four features: shared norms and values among community members, higher transparency and interactivity, lower barriers to participation, and strengthened collaboration between the platform and users.

However, tensions do exist. While the participatory model breaks down the traditional divide between platform-driven and user-driven moderation by adopting a cooperative mechanism, it introduces complexities in the power dynamics between users and platforms. On the one hand, it offers users a more active role in content moderation and community building. On the other hand, it raises concerns, such as the platform setting

the rules and leveraging users' voluntary efforts, mismatches between user expertise and the cases they handle, and the exclusion of politically sensitive content from the user moderation process, which limits the system's inclusivity and democratic potential. This narrow scope restricts opportunities for users to engage with more controversial or politically charged content, which would be essential for true democratic discourse.

Furthermore, profit-driven platforms like Bilibili can leverage user participation in this mechanism for enhancing content moderation based on user acceptance, rather than pursuing fairness and justice. Although user acceptance and justice are not mutually exclusive, the prioritization of user acceptance may lead to content moderation decisions that align more closely with the preferences of the majority, rather than ensuring equitable treatment for all users and minority voices when dealing with complicated cases involving ideological controversies.

While the system marks a shift towards more cooperative governance and offers a more transparent, accessible form of content moderation, it does not fully guarantee justice or democracy in moderation decisions. The platform's ability to manage the scope of user participation means that user engagement is largely confined to non-sensitive, low-conflict content that aligns with the platform's broader objectives and national regulations. This undermines the potential of participatory content moderation to foster democratic dialogue and mitigate the risks of content censorship in the Chinese context.

This research provides a unique insider perspective on Bilibili's participatory content moderation system, contributing to a deeper understanding of the challenges and potential of user-participatory content moderation in the context of Chinese platform governance. While the study focuses on platform-centric designs, future research could expand on users' perspectives and explore how different user motivations and struggles shape their involvement in content moderation. Moreover, further investigation into the broader implications of this model for global digital governance is necessary to understand its potential as a model for other platforms facing similar issues with user-generated content.

Notes

1. Weibo: China's leading microblogging platform.
   Douyin: A short-video sharing platform in China, it's the domestic version of TikTok and both are owned by ByteDance.
   Zhihu: The biggest Q&A platform in China.
   Meituan: A shopping platform provides food delivery and local services
2. Bilibili coins is the free virtual currency used on Bilibili. Users can earn free coins by using the app every day, publishing videos, and joining the disciplinary committee.

Author contributions

CRediT: **Zheyu Shang:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Zheyu Shang is a PhD candidate at the Institute for Area Studies, Leiden University. Her doctoral project focuses on the evolvement of China's online propaganda strategies in a platform age. Her research interests include platform governance, political communication, and youth culture in mainland China.

ORCID

Zheyu Shang http://orcid.org/0000-0002-6135-5628

References

Bilibili. (2021). B站百亿&年度弹幕盛典 [Bilibili Ceremony of 10 Billion Danmu & the Annual Top 1 Danmu Comment]. Retrieved November 8, 2022, from www.bilibili.com website: https://www.bilibili.com/blackboard/topic/activity-ygXy0F1OGW.html

Bilibili. (2022). 2021 *annual report on form 20-F*. Bilibili. https://ir.bilibili.com/media/55fpkkjg/2021-annual-report-on-form-20-f.pdf

Bilibili. (2023). 2022 *annual report on form 20-F*. https://ir.bilibili.com/media/rwafkhml/annual-and-transition-report-of-foreign-private-issuers-sections-13-or-15-d.pdf

Bilibili. (n.d.). *社区规范 [Community Guidelines]*. Bilibili.com. Retrieved February 9, 2024, from https://www.bilibili.com/blackboard/blackroom.html?spm_id_from = 888.20498.b_4d7144577a6c622d7355.3

Caplan, R. (2018). *Content or Context Moderation?* https://apo.org.au/sites/default/files/resource-files/2018-11/apo-nid203666.pdf

Caplan, R. (2023). Networked platform governance: The construction of the democratic platform. *International Journal of Communication*, *17*(22), 3451–3472.

Chen, Z. T. (2018). Poetic prosumption of animation, comic, game and novel in a post-socialist China: A case of a popular video-sharing social media Bilibili as heterotopia. *Journal of Consumer Culture*, *21*(2), 257–277. https://doi.org/10.1177/1469540518787574

Chen, Z. T. (2020). Slice of life in a live and wired masquerade: Playful prosumption as identity work and performance in an identity college Bilibili. *Global Media and China*, *5*(3), 319–337. https://doi.org/10.1177/2059436420952026

Chen, Y., Gao, Q., & Rau, P.-L. P. (2017). Watching a movie alone yet together: Understanding reasons for watching Danmaku Videos. *International Journal of Human–Computer Interaction*, *33*(9), 731–743. https://doi.org/10.1080/10447318.2017.1282187

Chen, X., Kaye, D. B. V., & Zeng, J. (2021). #PositiveEnergy Douyin: Constructing "playful patriotism" in a Chinese short-video application. *Chinese Journal of Communication*, *14*(1), 97–117. https://doi.org/10.1080/17544750.2020.1761848

Chen, Z., & Yang, D. L. (2023). Governing Generation Z in China: Bilibili, bidirectional mediation, and online community governance. *The Information Society*, *39*(1), 1–16. https://doi.org/10.1080/01972243.2022.2137866

Clark, P. (2012). *Youth culture in China: from Red Guards to netizens*. Cambridge Univ. Press.

Cook, C. L., Patel, A., & Wohn, D. Y. (2021). Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Frontiers in Human Dynamics*, *3*, 1–8. https://doi.org/10.3389/fhumd.2021.626409

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. https://doi.org/10.1177/1461444814543163

Dias Oliva, T., Antonialli, D. M., & Gomes, A. (2020). Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online. *Sexuality & Culture*, *25*(2), 700–732. https://doi.org/10.1007/s12119-020-09790-w

Dieter, M., Gerlitz, C., Helmond, A., Tkacz, N., van der Vlist, F. N., & Weltevrede, E. (2019). Multi-situated app studies: Methods and propositions. *Social Media + Society*, *5*(2), https://doi.org/10.1177/2056305119846486

Duguay, S., & Gold-Apel, H. (2023). Stumbling blocks and alternative paths: Reconsidering the walkthrough method for analyzing apps. *Social Media + Society*, *9*(1), https://doi.org/10.1177/20563051231158822

Eldon, E. (2010, January 4). Facebook begins testing advanced crowd-sourced content moderation. Retrieved August 23, 2023, from www.adweek.com website: https://www.adweek.com/performance-marketing/facebook-begins-testing-advanced-crowd-sourced-content-moderation/

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720943234

Gillespie, T., Aufderheide, P., Carmi, E., Gerrard, Y., Gorwa, R., Matamoros-Fernández, A., … Myers West, S. (2020). Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, *9*(4), 1–29. https://doi.org/10.14763/2020.4.1512

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–35. https://doi.org/10.1145/3479610

Han, S. (2016). 弹幕视频与参与式文化的新特征 [New Features of Participatory Culture and Videos with Danmu]. *新闻界*, (22), 54–57.

He, T. (2022). 观看"作为再创作：论视听文化再生产与受众介入式审美——基于技术可供性的视角 [Viewing as a form of reproduction: Audiovisual cultural reproduction and the audience's intrusive aesthetic - from a perspective of technology affordance]. *Modern Communication*, 125–132. https://doi.org/10.19997/j.cnki.xdcb.2022.04.009

Helberger, N., Pierson, J., & Poell, T. (2018). Governing online platforms: From contested to cooperative responsibility. *The Information Society*, *34*(1), 1–14. https://doi.org/10.1080/01972243.2017.1391913

Jenkins, H. (2009). *Confronting the challenges of participatory culture: Media education for the 21st century*. MIT Press.

Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable media: Creating value and meaning in a networked culture* (pp. 1–46). New York University Press.

Jenkins, H., Ito, M., & Boyd, D. (2015). *Participatory culture in a networked era: A conversation on youth, learning, commerce, and politics*. Polity Press.

Kaye, D. B. V., Chen, X., & Zeng, J. (2021). The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok. *Mobile Media & Communication*, *9*(2), 229–253. https://doi.org/10.1177/2050157920952120

Light, B., Burgess, J., & Duguay, S. (2018). The walkthrough method: An approach to the study of apps. *New Media & Society*, *20*(3), 881–900. https://doi.org/10.1177/1461444816675438

Manovich. (2001). *The language of new media*. Mit Press.

Mao, W., & Liao, S. (2020). 会员参与平台治理对用户黏性的影响 ———基于 BILIBILI的单案例研究 [The influence of users' participation in platform governance on user loyalty – a case study of Bilibili]. 管理案例研究与评论 *[Journal of Management Case Studies]*, *13*(1), 71–85. https://doi.org/11.7511/JMCS20200105

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit Blackout. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1138–1151. https://doi.org/10.1145/2858036.2858391

Matias, J. N. (2019). The civic labor of volunteer moderators online. *Social Media + Society*, *5*(2), https://doi.org/10.1177/2056305119836778

Mejias, U. A., & Couldry, N. (2019). Datafication. *Internet Policy Review*, *8*(4), 1–10. https://doi.org/10.14763/2019.4.1428

Miller, D., Costa, E., Haynes, N., McDonald, T., Nicolescu, R., Sinanan, J., … Wang, X. (2016). *How the world changed social media*. UCL Press.
Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4), 1–12. https://doi.org/10.1093/jcmc/zmac010
Nieborg, D. B., & Poell, T. (2018). The platformization of cultural production: Theorizing the contingent cultural commodity. *New Media & Society*, *20*(11), 4275–4292. https://doi.org/10.1177/1461444818769694
Oh, D., & Downey, J. (2024). Does algorithmic content moderation promote democratic discourse? Radical democratic critique of toxic language AI. *Information Communication & Society*, 1–20. https://doi.org/10.1080/1369118x.2024.2346531
Poell, T., Nieborg, D. B., & Brooke Erin, D. (2021). *Platforms and cultural production*. Polity Press.
Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4), 1–13. https://doi.org/10.14763/2019.4.1425
Pradel, F., Zilinsky, J., Kosmidis, S., & Theocharis, Y. (2024). Toxic speech and limited demand for content moderation on social media. *American Political Science Review*, 1–18. https://doi.org/10.1017/s000305542300134x
Rheingold, H. (2008). Using participatory media and public voice to encourage civic engagement. In W. Lance Bennett (Ed.), *Civic life online: Learning how digital media can engage youth* (pp. 97–118). The MIT Press. https://doi.org/10.1162/dmal.9780262524827.097
Roberts, S. T. (2019). *Behind the screen*. Yale University Press.
Schneider, F. (2021). China's viral villages: Digital nationalism and the COVID-19 crisis on online video-sharing platform Bilibili. *Communication and the Public*, *6*(1-4), 48–66. https://doi.org/10.1177/20570473211048029
Seering, J. (2020). Reconsidering self-moderation: the role of research in supporting community-based models for online content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–28. https://doi.org/10.1145/3415178
Suzor, N. P., West, S. M., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 1526–1543.
Thach, H., Mayworm, S., Delmonaco, D., & Haimson, O. (2022). (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society*, 1–22. https://doi.org/10.1177/14614448221109804
Twitter. (n.d.). About community notes on Twitter | Twitter Help. Retrieved from help.twitter.-comwebsite: https://help.twitter.com/en/using-twitter/community-notes
van Dijck, J. (2009). Users like you? Theorizing agency in user-generated content. *Media, Culture & Society*, *31*(1), 41–58. https://doi.org/10.1177/0163443708098245
van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society. Public values in a connective world*. Oxford University Press.
Wang, R. (2022). Community-building on Bilibili: The social impact of danmu comments. *Media and Communication*, *10*(2), 54–65. https://doi.org/10.17645/mac.v10i2.4996
Wu, Q., Sang, Y., & Huang, Y. (2019). Danmaku: A new paradigm of social interaction via online videos. *ACM Transactions on Social Computing*, *2*(2), 1–24. https://doi.org/10.1145/3329485
Xiang, Y., & Chae, S. W. (2021). Influence of perceived interactivity on continuous use intentions on the danmaku video sharing platform: Belongingness perspective. *International Journal of Human–Computer Interaction*, 1–21. https://doi.org/10.1080/10447318.2021.1952803
Yang, Y. (2020). The danmaku interface on Bilibili and the recontextualised translation practice: a semiotic technology perspective. *Social Semiotics*, *30*(2), 254–273. https://doi.org/10.1080/10350330.2019.1630962
Yin, Y., & Fung, A. (2017). Youth online cultural participation and Bilibili. In L. Rocci & R. Baarda (Eds.), *Digital media integration for participatory democracy* (pp. 130–154). IGI Global. https://doi.org/10.4018/978-1-5225-2463-2.ch007

Zeng, J., Kaye, D., & Bondy, V (2022). From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet*, *14*(1), 79–95. https://doi.org/10.1002/poi3.287

Zhang, W., & Mao, C. (2013). Fan activism sustained and challenged: Participatory culture in Chinese online translation communities. *Chinese Journal of Communication*, *6*(1), 45–61. https://doi.org/10.1080/17544750.2013.753499

Zhao, Y., Zhang, Z., Yao, J., & Zhou, Z. (2019). 社交平台自治组织的治安功能及其治理模式 [The public security function and governance model of social platform's autonomous organizations]. *贵州警察学院学报 [Journal of Guizhou Police College]*, *32*(2), 87–94. https://doi.org/10.13310/j.cnki.gzjy.2020.02.012

Zheng, X. (2019). 蔡徐坤出圈记之B站篇：律师函与网友的狂欢—36氪 [Cai Xukun"s journey to wider recognition on Bilibili: Legal letters and netizens" revelry. – 36Kr]. Retrieved August 22, 2023, from 36kr.com website: https://36kr.com/p/1723522514945

Zimbardo, P. G. (1969). The human choice: Individuation, reason and order versus deindividuation, impulse and chaos. In W. J. Arnold & D. Levine (Eds.), *Nebraska symposium on motivation* (pp. 237–307).