ENTREPRENEURSHIP AND SUSTAINABILITY ISSUES
ISSN 2345-0282 (online) http://jssidoi.org/jesi/
2025 Volume 12 Number 4 (June)
http://doi.org/10.9770/j8796966738

Publisher
http://jssidoi.org/esc/home

A COMPARATIVE ANALYSIS OF SOCIAL NETWORKS CONTENT MODERATION GUIDELINES AND BLOG POSTS[*]

Gintarė Gulevičiūtė [1], Monika Mačiulienė [2], Aelita Skaržauskienė [3], Asta Zelenkauskaitė [4], Aistė Diržytė[5]

Mykolas Romeris University, Ateities Street 20, Vilnius, Lithuania

E-mails: gintare.guleviciute@mruni.eu [1]; maciuliene@mruni.eu [2]; aelita@mruni.eu [3]; az358@drexel.edu [4]; aiste.dirzyte@mruni.eu [5]

Received 10 December 2025; accepted 25 March 2025; published 30 June 2025

Abstract. The study aims to examine the content moderation guidelines of major social media platforms and assess how these platforms communicate their moderation practices through official blog posts. A comparative analysis of Instagram, Facebook, TikTok, Snapchat, LinkedIn and X (formerly Twitter) revealed that while platforms share common moderation goals—such as prohibiting hate speech, misinformation, and harmful content—there are notable differences in the emphasis of these issues in public communications. Blog posts, used as self-representation tools, often focus on specific platform priorities, such as privacy or community protection, which may not always align with the full scope of the guidelines. For example, while misinformation is a formal concern across platforms, it is not always as prominently featured in public messaging. These findings highlight a gap between formal content moderation policies and how platforms publicly communicate their enforcement efforts.

Keywords: social networks; content moderation; blog posts

Reference to this paper should be made as follows: Gulevičiūtė, Mačiulienė, Skaržauskienė, Zelenkauskaitė, Diržytė, 2025. A comparative analysis of social networks content moderation guidelines and blog posts. *Entrepreneurship and Sustainability Issues*, 12(4), 118-128. http://doi.org/10.9770/j8796966738

JEL Classifications: M30, M10, D83

Additional disciplines: Management, Information and Communication

1. Introduction

The rapid rise of social media platforms has transformed communication and content-sharing dynamics within society (Wang et al., 2023). This transformation has presented new challenges for platforms, especially regarding harmful content such as misinformation. As platforms evolve, they face increased scrutiny to control the dissemination of harmful content through implementing content moderation systems (Keulenaar et al., 2023). These systems are formalized through community guidelines and policies aimed at regulating the flow of information (Singhal et al., 2023). However, while platforms outline strict guidelines, enforcement often falls short, allowing harmful content to persist (Gruzd et al., 2023; Arya et al., 2024; Shen et al., 2024). Content moderation systems can be broken down into distinct phases. Common (2020) outlines these as the creation of policies, the enforcement of these policies through detection and decision-making, and finally, the response to violations. This study will primarily focus on the creation phase—understanding the formal guidelines

[*] *This research was funded by the CHIST-ERA grant CHIST-ERA-21-OSNEM-004, by Research Council of Lithuania (Grant No. S-CHIST-ERA-22-1).*

established by major social media platforms and evaluating how they communicate these moderation strategies through public channels, such as blog posts.

Existing research highlights inconsistencies in how platforms moderate content and define guidelines (Singhal et al., 2023). While certain abusive content, such as hate speech and misinformation, is prohibited across platforms, there is no uniform approach to what constitutes a violation or how platforms address it. Furthermore, prior research focuses primarily on enforcement strategies (Myers West, 2018), discourse analysis (Gillespie, 2010) and labour involved in moderation (Roberts, 2016). However, there remains a need for a more comprehensive understanding of the formal content moderation guidelines and how they are communicated to users.

This **study aims** to address this gap by analyzing the content moderation guidelines of major social media platforms and examining the alignment between these formal policies and the communication of moderation practices through official blog posts.

Objectives of the research:

- To provide a comprehensive literature review on existing research regarding content moderation guidelines across social media platforms.
- To conduct an empirical comparison of the official blog posts of Instagram, Facebook, TikTok, Snapchat, LinkedIn, and X (formerly Twitter) with their stated content moderation guidelines.
- To identify gaps between the platform content moderation policies and their public communication via official blogposts.

2. Literature review

Content moderation on social media platforms such as Facebook, Instagram, TikTok, Snapchat and X (formerly Twitter) has become a crucial research topic due to the complexity of managing harmful content. Despite significant investments in content moderation technologies, platforms continue to face a variety of challenges. Gongane et al. (2022) highlight that these investments are often insufficient, as malicious content remains widespread. Platforms face difficulties detecting and moderating malicious content effectively, with content moderation models unable to handle the diversity and complexity of information across multiple modalities. This often leads to failures in detecting toxic content (Wang et al., 2023). Compounding the issue is the sheer volume of content generated daily, which makes it nearly impossible for human moderators or algorithms to catch everything. Udupa et al. (2023) and Liu et al. (2022) emphasize the scale of this challenge, pointing out that platforms are also pressured to balance user satisfaction and monetization, which can dilute their focus on moderation.

AI-driven content moderation systems have emerged as a popular solution for platforms, yet these systems are far from foolproof. Wang et al. (2023) and Valentine and Dupont (2022) argue that, while AI has the potential to process large amounts of data quickly, it still struggles to understand the nuances of human language and context, particularly when identifying hate speech or misinformation. These systems often fail to differentiate between harmful and benign content, leading to either over-enforcement or under-enforcement. Additionally, algorithmic biases within moderation tools pose another significant problem. Peterson-Salahuddin (2024) discusses how these biases can disproportionately affect marginalized groups, leading to further alienation of already vulnerable communities. Platforms have been criticized for silencing marginalized voices while failing to curb harmful content effectively, suggesting that the algorithms in place are not as effective as they could be (Haimson et al., 2021).

Each social media platform has developed its approach to content moderation, leading to platform-specific challenges and strategies. For instance, Keulenaar et al. (2023) focus on X, showing how its use of modulated moderation aims to create a crisis-resistant speech architecture. This strategy allows flexible content management during high-pressure situations, such as political elections or public health crises. However, these strategies often come under scrutiny, particularly when moderation efforts fail to curb the spread of harmful content in real time. TikTok faces unique challenges in moderating its predominantly video-based content.

Lookingbill and Le (2024) highlight how TikTok sometimes marginalizes user-generated content in an attempt to maintain community standards. Yet users have found ways to circumvent moderation algorithms by changing keywords or inventing new slang to bypass restrictions, as discussed by Steen et al. (2023). In addition to technological and platform-specific challenges, the ethical implications of content moderation have become a prominent issue in recent research. Franco et al. (2023) emphasize that content moderation systems, while essential, can have unintended consequences on user engagement, particularly when they disproportionately affect fragile and marginalized users.

3. Research methodology

This research utilized a comparative analysis to evaluate major social media platforms' content moderation guidelines and practices (Ragin, 1987; Bennett & George, 2005). Data was collected from official blog posts of Instagram, Facebook, TikTok, Snapchat, LinkedIn, and X (formerly Twitter), covering the period from 2018 to 2024 (with X data extending up to 2022). Comparative analysis enhances decision accuracy and helps generate broader social theories from specific case studies (Przeworski & Teune, 1970). This approach facilitates a deeper understanding of misinformation management and platform accountability. To identify patterns and discrepancies, the study analyzed both content moderation guidelines and blog posts, focusing on themes related to misinformation and harmful content. AntConc, a freeware corpus analysis toolkit, was employed to mine text and identify key themes in blog communications. As blogs serve as a self-representation tool, their analysis offers valuable insights into how platforms publicly discuss and highlight their moderation practices (Gilpin, 2010; Buckley & Schafer, 2022), revealing not only enforcement practices but also evolving platform priorities.

Case selection

The selection of top social media platforms was based on their global popularity and user engagement, which are crucial factors for assessing the effectiveness of content moderation guidelines. The platforms chosen—Instagram, Facebook, TikTok, Snapchat, LinkedIn, and X (formerly Twitter)—represent a broad spectrum of target audiences and content types, allowing for a comprehensive analysis of diverse moderation practices. The selected platforms rank among the highest in terms of daily active users, global reach and content volume (Data Reportal, 2024). The platforms cater to different demographics, from professional networks (LinkedIn) to youth-oriented platforms (TikTok), ensuring that the analysis covers a wide range of moderation challenges.

Data collection

Data was gathered from two primary sources: content moderation guidelines and official blog posts on Instagram, Facebook, TikTok, Snapchat, LinkedIn, and X (formerly Twitter). The current guidelines were scraped from the internet, with guideline lengths varying from 3,000 to 10,000 characters depending on the platform. These guidelines were collected and categorized by themes such as misinformation, hate speech, privacy, and harassment. Additionally, official blog posts published by Instagram (https://about.instagram.com/blog/), Facebook (https://about.fb.com/news/), TikTok (https://newsroom.tiktok.com/en-us), Snapchat (https://newsroom.snap.com/safety-and-impact), LinkedIn (https://www.linkedin.com/blog/member) and X (https://blog.x.com/en_us) between 2018 and 2024 were collected, with X posts analyzed up to 2022. Tech company blogs primarily serve to communicate new policy decisions or initiatives, particularly those aimed at addressing harmful behaviours on their platforms. While the content varies, these blogs typically explain the rationale behind policy changes and emphasize the platforms' efforts to enhance user safety and counteract malicious activities (Watkin & Conway, 2022). All posts were manually reviewed, and only those relevant to the study’s focus on content moderation practices were selected. The blog posts were organized into a spreadsheet, documenting the social network, blog title, publication date, URL, and the full text of each relevant post. In total, 917 blog posts were analyzed, with the quantity of posts per platform shown in Table 1 below.

Table 1. The quantity of analyzed posts

Social network	Analyzed period	Relevant blog posts
Instagram	2018-2024	41
Facebook	2018-2024	541
TikTok	2018-2024	126
Snapchat	2018-2024	112
LinkedIn	2018-2024	28
X (previous Twitter)	2018-2022	69
Total:		**917**

Data analysis

First, the data collected on moderation guidelines and blogposts were analyzed using a qualitative approach through Nvivo, a tool designed for managing and analyzing qualitative data. A bottom-up approach was employed, where inductive codes were created and then grouped into broader themes. Two coders independently coded the data, and their outputs were compared. Discrepancies were resolved by consensus, ensuring consistency in coding. Following the thematic analysis, a quantitative comparison of blog posts and guidelines was carried out, where the frequency of key themes—such as safe usage, community protection, and privacy—was counted and compared across platforms. To further analyze blog post content, the AntConc (https://www.laurenceanthony.net/software/antconc/) tool was used to create wordlists and identify keywords, allowing for the analysis of the context in which key terms were used. This process helped to refine the thematic findings and ensured comprehensive coverage of the moderation topics.

Limitations

Any interpretative techniques in qualitative analysis can introduce subjectivity, even with efforts to mitigate bias through coder comparison. The reliance on official blog posts and public guidelines may not fully reflect real-world enforcement practices, as platforms may curate their communication. Additionally, the focus on data from 2018-2024 limits the scope, potentially overlooking longer-term trends or recent developments in content moderation strategies.

5. Findings

Results of content moderation guidelines analysis

The summary of the key themes in content moderation guidelines across social media platforms is provided in Table 2. The analysis highlights that the core moderation policies are largely consistent across Instagram, Facebook, TikTok, Snapchat, LinkedIn and X (formerly Twitter) despite differences in platform focus and user demographics. The platforms in the sample universally prohibit hate speech, which is defined as content promoting violence or discrimination based on personal characteristics such as race, gender, religion, ethnicity, sexual orientation or disabilities. Similarly, all platforms enforce strict rules against bullying, harassment, and threats, ensuring a safer online environment.

Table 2. Findings of moderation guidelines analysis

What is not allowed according to content moderation guidelines	**Social networks**					
	Instagram	*Facebook*	*Snapchat*	*LinkedIn*	*TikTok*	*X (Twitter)*
Harmful content and new types of abuse related to COVID-19	x					
Hate speech	x	x	x	x	x	x
Coordination of harm	x	x	x		x	
Bullying and harassment	x	x	x	x	x	x
Misinformation	x		x	x	x	x
Nudity	x	x	x		x	
Artificially collecting likes, followers, or shares, posting repetitive comments or content, or repeatedly contacting people for commercial purposes without their consent.	x	x		x	x	x
Impersonate others	x	x		x	x	x
Support or praise terrorism, organized crime, or hate groups.	x	x	x	x	x	x
Offering sexual services, buying or selling firearms, alcohol, and tobacco products, drugs	x	x	x	x	x	x
Sharing sexual content involving minors or threatening to post intimate images of others.	x	x	x	x	x	x
Serious threats of harm to public and personal safety.	x	x	x	x		x
Encouraging or urging people to embrace self-injury, suicide.	x	x			x	x
Intense, graphic violence in videos.	x	x		x	x	x
Human Exploitation		x	x			
Privacy Violations		x				x
Attempts to gather sensitive user information or engage in unauthorized access		x	x	x		x
Violation of Intellectual Property		x			x	x
Share harmful or shocking material				x		
Grooming behaviours					x	
Civic Integrity						x
Third-party advertising in video content						x

A significant emphasis is placed on protecting minors, with all platforms adopting a zero-tolerance policy toward child exploitation, including child sexual abuse, grooming, and other harmful activities related to minors. Furthermore, there are clear prohibitions on the sharing of explicit sexual content, particularly non-consensual nudity and child sexual abuse material. In terms of violence, each platform bans content that incites violence or encourages self-harm or suicide. Support for terrorism, hate groups, and organized crime is strictly forbidden across the board. Lastly, user privacy is a critical concern, with platforms prohibiting sharing personal data without consent and cracking down on privacy violations.

Despite these common policies, each platform demonstrates its unique priorities depending on its user base and content focus. While there is a shared commitment to addressing major issues such as hate speech, violence, harassment, and child safety, the scope and focus of content moderation vary significantly according to the nature of each platform. For example, Instagram and TikTok emphasize visual content moderation. Instagram has strict rules concerning graphic violence, nudity, and harmful visual trends, as evidenced by their guideline: *"Post only your own photos and videos and always follow the law. Respect everyone on Instagram, don't spam people or post nudity"* (Instagram, September 2024). Similarly, TikTok applies strong moderation to prevent harmful visual trends, stating: *"To minimize the potentially negative impact of graphic content, we may first include safety measures such as an 'opt-in' screen or warning"* (TikTok, September 2024). This is particularly important on TikTok, given its large younger audience and its focus on short, viral video content.

On the other hand, LinkedIn emphasizes the importance of maintaining a professional environment, which shapes its approach to content moderation. Its guidelines disallow content that deviates from work-related or ethical standards, promoting a professional and respectful space: *"Only bring safe conversations to LinkedIn. We allow broad conversations about the world of work, but just keep it professional"* (LinkedIn, September 2024). Finally, X (formerly Twitter), given its role as a platform for public discourse, centres its moderation guidelines around regulating abusive speech, civic integrity, and the manipulation of public information. The platform's rules state: *"X's purpose is to serve the public conversation. Violence, harassment, and similar types of behaviour discourage people from expressing themselves and ultimately diminish the value of global public conversation. Our rules are to ensure all people can participate freely and safely"* (X, September 2024).

This analysis underscores the shared responsibilities among platforms to tackle harmful content while highlighting their platform-specific challenges. The differences in content moderation approaches reflect the platforms' varying priorities, influenced by the type of content they host and the audience they serve.

Results of blogpost analysis

Official blog posts provide a narrative directly from the platforms, frequently detailing updates to policies, enforcement strategies and the platforms' positions on critical issues such as misinformation. Analyzing these posts allowed us to understand better how platforms communicate their moderation efforts and the role this communication plays in shaping user behaviour and the spread of misinformation. The analysis using AntConc identified key topics across the social media platforms studied, focusing on themes such as safe usage, community protection, privacy, and misinformation (see Table 3).

Table 3. The main topics of analyzed blog posts

Social network and main topics	Relevant blog posts	Word frequencies
Instagram		
Community protection	19	151 (0.2%)
Safe usage	13	124 (0.2%)
COVID-19, antibullying, hate speech, misinformation	9	94 (0.2%)
Facebook		
Data and privacy	230	1324 (0.7%)
Safety and expression	224	1285 (0.7%)
Combating misinformation	87	356 (0.2%)
Tik Tok		
Important news	13	104 (0.6%)
Community	11	92 (0.4%)
Safe usage	102	315 (0.2%)
Snapchat		
Community	22	116 (0.6%)
Parental tools	18	145 (0.6%)
Safe usage	72	384 (0.7%)
LinkedIn		
Safe usage	14	294 (0.3%)
Updates to community policies	6	94 (0.2%)
Transparency	8	118 (0.2%)
X (previous Twitter)		
Safety	36	224 (0.7%)
Transparency	33	284 (0.4%)

The blog posts were particularly insightful in revealing how platforms articulate their commitment to safety and moderation while showcasing differences in priorities based on platform demographics and content types. For instance, Instagram primarily focused on creating a safe online environment, with a strong emphasis on community protection and safe usage. Its blog posts frequently addressed efforts to foster a healthy and inclusive space, particularly in response to challenges such as misinformation during the COVID-19 pandemic. The posts highlighted various initiatives to maintain a safe community and combat harmful content trends. Facebook, on the other hand, dedicated significant attention to data privacy and user safety, likely in response to its large user base and its history of privacy-related controversies. The high frequency of posts regarding privacy and safety suggests the platform is actively engaging with these issues. However, while misinformation is a critical topic, its relatively low word frequency (0.2%) in blog posts implies it receives less focus compared to privacy and safety concerns.

The analysis showed that TikTok, given its younger demographic, emphasized safe usage to address potential risks of harmful content. Blog posts highlighted topics such as "Important News" (0.6% frequency), reflecting efforts to combat misinformation and potentially dangerous trends like harmful challenges or misleading content that could endanger users. Similarly, Snapchat, with its younger user base, focused extensively on parental tools and community protection. The high frequency of posts related to safe usage (0.7%) suggests that ensuring safe interactions, particularly with parental supervision tools, is a priority for the platform. LinkedIn's blog posts were centred around maintaining a professional and respectful environment, with a particular focus on safe usage and transparency. The lower volume of discussions compared to other platforms likely reflects its professional nature, where content moderation aims to uphold professionalism rather than mitigate a wide array of content types. Finally, X (formerly Twitter) dedicated a significant portion of its blog posts to user safety and transparency, particularly focusing on improving user trust during its transformation. The relatively high frequency of keywords in these posts reflects the platform's efforts to position itself as a leader in public discourse moderation and misinformation management.

Gaps between content moderation guidelines and public communication

To analyze the gap between formal content moderation policies and how they are represented or discussed in platforms' public communications, we can identify several key themes by comparing the moderation guidelines (Table 2) with the focus of official blog posts (Table 3). Few discrepancies have been identified in this regard. First, the content moderation guidelines across platforms like Facebook, Instagram, and TikTok clearly prohibit misinformation, particularly on sensitive topics like health (e.g., COVID-19) and civic processes (e.g., elections). This commitment is outlined in their guidelines (Table 2). However, the frequency of blog posts discussing "combating misinformation" is notably lower. For instance, on Facebook, only 87 blog posts (0.2% of the total) mention combating misinformation (Table 3). This discrepancy suggests that, although the platforms have robust guidelines on misinformation, it is not prominently featured in their public communications. This gap between strict policies and their public emphasis may reflect a strategic focus on other issues in their outward communication, potentially diluting their stance on misinformation for broader audiences.

Second, the platforms universally emphasize the need to protect users from harmful content like hate speech, harassment, and violence, as seen in their moderation guidelines (Table 2). However, in their public communications, platforms like X (formerly Twitter) and Facebook tend to focus more on transparency, data privacy, and community safety, as shown by the higher frequency of blog posts discussing these topics. For example, X mentions user safety 36 times in blog posts, while direct discussions on harmful content, such as hate speech, are less frequent. This shift in focus suggests that while these platforms are committed to user safety in their formal policies, the narrative in public communication may prioritize other concerns, such as privacy or transparency, possibly to align with public discourse or regulatory pressures.

Finally, although the content moderation policies are consistent in addressing key issues like misinformation and harmful content, each platform tailors its public communication based on its unique user base. Instagram and TikTok emphasize community protection and visual content moderation in their blog posts, aligning with the visual nature of their platforms. For instance, Instagram’s guideline to "Post only your own photos and videos and follow the law" (Instagram, September 2024) underscores the platform’s focus on visual content integrity. In contrast, LinkedIn emphasizes professionalism both in its moderation guidelines and public communications, reflecting its goal to maintain a respectful and professional environment. This variation shows that while moderation policies may be similar across platforms, the communication strategies vary to reflect platform-specific concerns and user demographics.

The results suggest that while content moderation guidelines are extensive and address a wide range of harmful content, the way platforms communicate these policies publicly, especially through blog posts, reveals gaps. Certain issues, like misinformation or harmful content, may not be as frequently highlighted in public communications as their policies would suggest. This discrepancy highlights a potential gap between the creation of formal content moderation policies and the platforms' public communication strategies, potentially shaped by the need to balance transparency, user trust and regulatory demands.

Discussion and conclusions

After analysis of the existing content moderation guidelines across major social media platforms, the results showed that platforms such as Instagram, Facebook, TikTok, Snapchat, LinkedIn, and X (formerly Twitter) share core moderation policies (bans on hate speech, misinformation, and harmful content), but each platform emphasizes different aspects based on its target audience and operational model. The findings from this research highlight significant insights into how social media platforms formulate, communicate and (to some extent) implement content moderation policies. Despite the existence of clear and robust moderation guidelines across major platforms, there remain critical discrepancies between these formal policies and the themes emphasized in their public communications, particularly in official blog posts.

The analysis of official social media blog posts showed that official blog posts serve as a strategic self-representation tool. The social media platforms selectively emphasize certain aspects of their content moderation efforts. For example, Facebook and LinkedIn focus on privacy and data security, while TikTok and Instagram highlight safe usage and community engagement. Although explicitly prohibited in guidelines, misinformation receives limited attention in blog posts.

One of the most notable gaps identified is the disparity between platforms' stated commitment to combating misinformation and how prominently this issue is discussed publicly. While misinformation is prohibited under formal guidelines, platforms like Facebook and Instagram do not seem to emphasize this issue in their blog posts to the same degree. This gap could reflect a broader strategic focus by platforms on privacy, transparency, and other topics they deem more pressing or publicly favourable. As such, this creates a potential disconnect between what platforms say they are doing to mitigate harmful content and how they publicly present these efforts.

The implications of these findings are twofold. First, the discrepancies identified between formal content moderation policies and public communication suggest that platforms may prioritize certain issues over others in their outward messaging. This could have consequences for public trust, particularly if users perceive that platforms are not being transparent about their efforts to mitigate harmful content. Second, the variation in communication strategies underscores the need for more platform-specific studies when analyzing the efficacy of content moderation.

The novelty of this research is focused on uncovering the disconnect between social media content moderation policies and how these policies are publicly framed, contributing to discourse on digital governance and platform accountability. Also, by utilizing both thematic analysis and frequency-based keyword examination, the study provides a multi-faceted understanding of moderation communication strategies. This study contributes to the ongoing debate on digital platform governance by highlighting the need for increased transparency in content moderation.

The findings align with earlier studies (Singhal et al., 2023; Keulenaar et al., 2023), highlighting inconsistencies in content moderation enforcement. Unlike Franco et al. (2023), who focus on algorithmic challenges, this study emphasizes the role of blog posts in shaping public perception. The results contrast Peterson-Salahuddin (2024), who argues that moderation policies are disproportionately strict for marginalized communities. Instead, this research suggests that platforms strategically manage their public communication rather than solely enforcing biased policies.

References

AntConc. (2024). *A freeware corpus analysis toolkit for concordancing and text analysis.* Retrieved September 25 https://www.laurenceanthony.net/software/antconc/

Arya, P. *et al*. (2024) MSCMGTB: A Novel Approach for Multimodal Social Media Content Moderation Using Hybrid Graph Theory and Bio-Inspired Optimization. *IEEE Access*, vol. 12, pp. 73700-73718, https://doi.org/10.1109/ACCESS.2024.3400815

Bennett, A., & George, A. L. (2005). *Case Studies and Theory Development in the Social Sciences*. MIT Press.

Buckley, N., & Schafer, J. S. (2022). 'Censorship-free'platforms: Evaluating content moderation policies and practices of alternative social media. *Media and far-right*, Vol. 4: special issue 1. https://doi.org/10.21428/e3990ae6.483f18da

Bruckman, A., Curtis, P., Figallo, C., Laurel, B. (1994). Approaches to managing deviant behavior in virtual communities. *Conference companion on Human factors in computing systems*, pp. 183-184. https://doi.org/10.1145/259963.260231

Common, M. F. (2020). Fear the reaper: How content moderation rules are enforced on social media. *International Review of Law, Computers & Technology*, *34*(2), 126-152. https://doi.org/10.1080/13600869.2020.1733762

Data Reportal (2024). Global Social Media Statistics. Retrieved September 28 https://datareportal.com/social-media-users

Facebook. (2024). *Facebook blog posts*. Facebook news. Retrieved September 25 https://about.fb.com/news/

Foreman G. (2022). *The Ethical Journalist: Making Responsible Decisions in the Digital Age*. Wiley Blackwell.

Franco, M., Gaggi, O., Palazzi, C. (2023). Analyzing the Use of Large Language Models for Content Moderation with ChatGPT Examples. In Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks (OASIS '23). Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3599696.3612895

García, A., Martín, A., Huertas-Tato, J., Camacho, D. (2023). Countering malicious content moderation evasion in online social networks: Simulation and detection of word camouflage, *Applied Soft Computing*, 145. https://doi.org/10.1016/j.asoc.2023.110552

Gibson, A. D. (2023). What Teams Do: Exploring Volunteer Content Moderation Team Labor on Facebook. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186109

Gillespie T (2010) The politics of "platforms." New Media & Society 12(3): 347–364. https://doi.org/10.1177/14614448093427

Gilpin, D. (2010). Organizational Image Construction in a Fragmented Online Media Environment. *Journal of Public Relations Research*, *22*(3), 265–287. https://doi.org/10.1080/10627261003614393

Gongane, V.U., Munot, M.V. & Anuse, A.D. (2022). Detection and moderation of detrimental content on social media platforms: current status and future directions. *Soc. Netw. Anal. Min.* 12. https://doi.org/10.1007/s13278-022-00951-3

Gruzd, A., Soares, F. B., & Mai, P. (2023). Trust and Safety on Social Media: Understanding the Impact of Anti-Social Behavior and Misinformation on Content Moderation and Platform Governance. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231196878

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://dx.doi.org/10.7302/2632

Haythornthwaite, C. (2023). Moderation, Networks, and Anti-Social Behavior Online. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231196874

Hong, T., Tang, Z., Lu, M., Wang, Y., Wu, J., & Wijaya, D. (2023). Effects of #coronavirus content moderation on misinformation and anti-Asian hate on Instagram. *New Media & Society*, *0*(0). https://doi.org/10.1177/14614448231187529

Huber, G. P., & Power, D. J. (1985). Retrospective Reports of Strategic-Level Managers: Guidelines for Increasing Their Accuracy. *Strategic Management Journal*, 6(2), 171–180. https://doi.org/10.1002/smj.4250060205

Instagram. (2024). *Instagram content moderation guidelines*. Instagram Help Center. Retrieved September 20 https://help.instagram.com/477434105621119

Instagram. (2024). *Instagram blog posts*. Instagram Blog. Retrieved September 20 https://about.instagram.com/blog/

Jhaver, S., & Zhang, A. X. (2023). Do users want platform moderation or individual control? Examining the role of third-person effects and free speech support in shaping moderation preferences. *New Media & Society*, *0*(0). https://doi.org/10.1177/14614448231217993

Keulenaar, E., Magalhães, J., Ganesh, B. (2023). Modulating moderation: a history of objectionability in Twitter moderation practices, *Journal of Communication*, Volume 73, Issue 3, Pages 273–287, https://doi.org/10.1093/joc/jqad015

LinkedIn. (2024). *LinkedIn content moderation guidelines*. LinkedIn Legal. Retrieved September 20 https://www.linkedin.com/legal/professionalcommunity-policies

LinkedIn. (2024). *LinkedIn blog posts*. LinkedIn Blog. Retrieved September 20 https://www.linkedin.com/blog/member

Liu, Y., Yildirim, P., Zhang, J. (2022) Implications of Revenue Models and Technology for Content Moderation Strategies. *Marketing Science* 41(4):831-847. https://doi.org/10.1287/mksc.2022.1361

Lookingbill, V., & Le, K. (2024). "There's Always a Way to Get Around the Guidelines": Nonsuicidal Self-Injury and Content Moderation on TikTok. *Social Media + Society*, *10*(2). https://doi.org/10.1177/20563051241254371

Meta. (2024). *Facebook content moderation guidelines*. Meta Transparency Center. Retrieved September 20 https://transparency.meta.com/policies/community-standards/

Meta. (2024). *Facebook blog posts*. Meta Newsroom. Retrieved September 20 https://about.fb.com/news/

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366-4383. https://doi.org/10.1177/1461444818773059

Ozanne, M., Bhandari, A., Bazarova, N. N., & DiFranzo, D. (2022). Shall AI moderators be made visible? Perception of accountability and trust in moderation systems on social media platforms. *Big Data & Society*, *9*(2). https://doi.org/10.1177/20539517221115666

Peterson-Salahuddin, C. (2024). Repairing the harm: Toward an algorithmic reparations approach to hate speech content moderation. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241245333

Przeworski, A., & Teune, H. (1970). *The Logic of Comparative Social Inquiry*. Wiley-Interscience.

Ragin, C. C. (1987). *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*. University of California Press.

Roberts, S. (2016) Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. Wi: *Journal of Mobile Media, 10*(1), 1–18.

Shen, Y., Wu, W., Cai, F., Song, L., & Zhang, Y. (2024). Influence Mechanism of Social Support of Social Media on Users' Information-Seeking Behavior. *Transformations in Business & Economics*, Vol. *23*, No 1(61), pp.443–469.

Singhal, M. et al. (2023). SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. *IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 868-895, https://doi.org/10.1109/EuroSP57164.2023.00056

Snap Inc. (2024). *Snapchat content moderation guidelines*. Snap Inc. Values. Retrieved September 20 https://values.snap.com/privacy/transparency/community-guidelines

Snap Inc. (2024). *Snapchat blog posts*. Snap Inc. Newsroom. Retrieved September 20 https://newsroom.snap.com/safety-and-impact

Steen, E., Yurechko, K., & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231194586

TikTok. (2024). *TikTok content moderation guidelines*. TikTok Community Guidelines. Retrieved September 20 https://www.tiktok.com/community-guidelines/en?lang=en

TikTok. (2024). *TikTok blog posts*. TikTok Newsroom. Retrieved September 20 https://newsroom.tiktok.com/en-us

Udupa, S., Maronikolakis, A., & Wisiorek, A. (2023). Ethical scaling for content moderation: Extreme speech and the (in)significance of artificial intelligence. *Big Data & Society*, *10*(1). https://doi.org/10.1177/20539517231172424

Valentine, C., & Dupont, B. (2022). Cognitive assemblages: The entangled nature of algorithmic content moderation. *Big Data & Society*, *9*(2) https://doi.org/10.1177/20539517221143361

Wang, W., Huang, J., Chen, C., Gu, J., Zhang, J., Wu, W., He, P., & Lyu, M. (2023). Validating Multimedia Content Moderation Software via Semantic Fusion. In Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2023). Association for Computing Machinery, New York, NY, USA, 576–588. https://doi.org/10.1145/3597926.3598079

Watkin, A. L., & Conway, M. (2022). Building social capital to counter polarization and extremism? A comparative analysis of tech platforms' official blog posts. *First Monday*. https://doi.org/10.5210/fm.v27i5.12611

X. (2024). *X content moderation guidelines*. X Help Center. Retrieved September 20 https://help.x.com/en/rules-and-policies/x-rules

X. (n.d.). *X blog posts*. X Blog. Retrieved September 20 https://blog.x.com/en_us

Funding: This research was funded by the CHIST-ERA grant CHIST-ERA-21-OSNEM-004, by Research Council of Lithuania (Grant No. S-CHIST-ERA-22-1).

Author Contributions: Conceptualization: *Gulevičiūtė, Mačiulienė, Skaržauskienė, Zelenkauskaitė, Diržytė*, methodology: *Gulevičiūtė, Mačiulienė, Skaržauskienė, Zelenkauskaitė, Diržytė*, data analysis: *Gulevičiūtė, Mačiulienė, Skaržauskienė,* writing—original draft preparation: *Zelenkauskaitė, Diržytė*, writing; review and editing: *Gulevičiūtė, Mačiulienė, Skaržauskienė, Zelenkauskaitė, Diržytė,* visualization: *Gulevičiūtė, Mačiulienė, Skaržauskienė, Zelenkauskaitė, Diržytė*. All authors have read and agreed to the published version of the manuscript.

Gintarė GULEVIČIŪTĖ is a researcher and lecturer at Mykolas Romeris university involved in a number of international research projects. Her areas of interest are digital marketing methods in digital platforms, e-commerce models for emerging markets, analysis of misinformation in online social networks and societal response to fake news together.
ORCID ID: https://orcid.org/0000-0003-1974-3982

Monika MAČIULIENĖ is a senior researcher at Mykolas Romeris University and associate professor and Vilnius Gediminas Technical University. Her research examines emergence of both collective intelligence and misinformation in social networks and their synergies and stakeholder engagement through co-creative measures. M. Mačiulienė actively participates in international conferences, seminars, and scientific exchange programs and published more than 40 research articles individually and in collaboration with international research teams. Currently, she is involved in a number of international scientific projects (e.g., Horizon Europe projects CLIMAS, DIGICHer, MARTINI H2020 projects INCENTIVE and EU-Citizen.Science) focused on synergies between science and the society as a senior researcher.
ORCID ID: https://orcid.org/0000-0002-8527-7468

Aelita SKARŽAUSKIENĖ is a professor at Mykolas Romeris University and chief researcher at Vilnius Gediminas Technical university. Her main research interests are digital co-creation, collective intelligence, decentralized Web and governance. Currently, she is leading the Lithuanian team in Horizon Europe projects CLIMAS and DIGICHer. Together with her research team, Aelita Skaržauskienė has developed a Collective Intelligence Monitoring Technique for evaluation of networked platforms (www.collective-intelligence.lt/en), in collaboration with the MIT Centre of Collective Intelligence.
ORCID ID: https://orcid.org/0000-0003-1606-0676

Asta ZELENKAUSKAITĖ research centers around emergent practices online by bridging multidisciplinary approaches drawn from social science tradition, Communication, Information Science and Linguistics. Her research focuses on the ways in which online interaction can create new spaces and practices for their users. Her work is interested in societal challenges of information mistrust and post-truth and the way such inauthentic information can be uncovered. She is focusing in the changes that social media bring to mass media landscape by studying these phenomena from a multi-method approach from macro and micro approaches.
ORCID ID: https://orcid.org/0000-0001-5762-4605

Aistė DIRŽYTĖ is a medical (clinical) psychologist, currently a professor at Vilnius Gediminas Technikos University and Mykolas Romeris University, and the Director of the Institute of Management and Psychology. Aistė is a full member of the American Psychological Association, a member of the International Association of Positive Psychology, a reviewer for the scientific journals Brain Sciences, Journal of Happiness Studies and many others, a reviewer for more than 50 scientific publications, co-author of two scientific studies, 5 textbooks. Recently, together with the Santaros Clinics, she has been analyzing information on the links between childhood traumatic experiences and mental health and testing the effectiveness of trauma-focused cognitive behavioral therapy interventions.
ORCID ID: https://orcid.org/0000-0003-2057-3108