Cambridge Forum on AI: Law and Governance (2025), **1**, e21, 1–19
doi:10.1017/cfl.2025.5

Cambridge Forum

RESEARCH ARTICLE

Meta's AI moderation and free speech: Ongoing challenges in the Global South

Soorya Balendra

Center for Human Rights and Legal Pluralism, McGill University, Montreal, Canada
Email: soorya.balendra@mail.mcgill.ca

(Received 30 October 2024; revised 2 March 2025; accepted 16 March 2025)

Abstract

This study investigates the discriminatory impact of artificial intelligence (AI)-driven content moderation on social media platforms (SMPs), particularly in the Global South, where cultural and linguistic diversity often clash with the Western-centric AI frameworks. Platforms like Meta increasingly rely on AI algorithms to moderate vast amounts of content, but research shows that these algorithms disproportionately restrict free expression in the Global South (European Union Agency for Fundamental Rights, 2023; De Gregorio & Stremlau, 2023). This results in "over removal" – censorship of lawful content – and "slow removal," which fails to address harmful material, both of which perpetuate inequality and hinder free speech. Through a case study on Meta, this research examines how AI-based content moderation misunderstands local contexts and systematically marginalizes users. The contributing factors include limited financial investment, inadequate language training, and political and corporate biases. The imbalance reflects power asymmetries, as governments in the Global South lack influence over platform policies. This study uses a human rights perspective to explore solutions through multistakeholder engagement, advocating for collaboration among tech companies, governments, and civil society to reform AI governance. Ultimately, it aims to inform regulatory frameworks that ensure fairer, more inclusive content moderation and protect free expression for a globally equitable digital landscape.

Keywords: Social Media Speech; Global South; AI Moderation; AI Bias; Meta

1. Introduction

Social media plays a crucial role in accommodating free speech in the digital space, facilitating political debate and communication, and offering accessible tools for users to organize and mobilize communities. More particularly, social media platforms (SMPs) provide spaces to produce and consume news that is no longer *elite-biased* (Ceron, 2015). Several studies highlight that social media has become a key alternative to traditional media outlets, such as television, radio, and newspapers. These traditional platforms are often criticized for promoting various forms of propaganda, bias, and deliberate misinformation, which pose significant challenges in protecting human rights and democratic values (Anderson, 2021; Elejalde et al., 2018).

Social media empower individuals, including those who have historically been excluded or marginalized (Ortiz et al., 2019; Schmitz et al., 2022), to connect with one another, access knowledge, culture, and information, raise awareness of human rights violations, and engage in political discourse in unprecedented ways (Garimella et al., 2018; Stieglitz & Dang-Xuan, 2013). Nevertheless,

these platforms also provide fertile ground for the proliferation of harmful content, including hate speech, intimidation, harassment, incitement to violence, and the spread of misinformation (San Martín, 2023). Recent events, including the Christchurch Mosque shootings, violence against the Rohingya population, and numerous disinformation campaigns related to the Brexit referendum and the 2016 U.S. elections, as well as Russia's invasion of Ukraine, exemplify the pressing need for regulatory measures addressing speech on social media (Bradford, 2023; Errington et al., 2020; Mudde, 2019). These incidents underscore the imperative to navigate the complexities surrounding speech in these open environments of social media – which cannot be treated as an *unfettered* right with minimal restrictions anymore.

The authority to regulate social media content is a topic of ongoing debate among government authorities and private stakeholders, including major tech companies. This discourse has led to the emergence of three principal regulatory models: (i) self-regulation, in which technology companies or the platforms bear the responsibility for moderating activities on their platforms; (ii) external regulation, wherein platform activities are subject to state oversight; and (iii) co-regulation, which represents a hybrid approach – with the participation of the State government authorities, and platforms to the establishment of rules and standards (Gillespie, 2017; Gorwa, 2019; Gosztonyi, 2023). In many jurisdictions, primary regulatory authority is vested in government entities, resulting in significant disparities in internet censorship standards worldwide. Granting absolute regulatory power to state actors often stifles political discourse and online expression, while the rise of authoritarian practices in the digital space emphasizes the importance of empowering platforms to take the lead in content moderation. The Chinese *Firewall* is one of the greatest examples of how an authoritarian approach suppresses freedom of speech and information flows across digital platforms.

Therefore, to mitigate the extensive content censorship prevalent today, contemporary movements increasingly advocate for a more significant role for private actors – specifically, the platforms – within the content regulation paradigm. Self-regulation has been adopted by many platforms, particularly by tech giants such as Meta and X, according to their content policies and community standards (De Abreu Duarte, 2024; Gorwa, 2024). These frameworks determine which content is permissible and which is prohibited, thereby influencing the dynamics of free speech and expression online (Chetty & Alathur, 2018; Klonick, 2017). Additionally, these platforms employ content moderation as a method to enforce their policies and standards (Gorwa et al., 2020; Myers West, 2018). Content moderation involves reviewing and managing user-generated content to ensure compliance with the established guidelines (Boberg et al., 2018; George & Scerri, 2007; Sander, 2019). However, due to the vast volume of content shared on these platforms, they increasingly rely on artificial intelligence (AI) for content moderation. This reliance introduces systematic biases that impact content removal decisions across platforms, leading to significant repercussions for core democratic values, which is discussed extensively in this paper.

This paper explores how the biases in Meta's AI-driven content moderation systems undermine free speech and expression, creating challenges for the public good in democratic societies. This study highlights how these biases, when applied in social media regulation, threaten free expression and compromise key democratic values, especially in the Global South. As an increasingly important market for platforms like Meta, the Global South offers a unique context for understanding the impact of AI-based content moderation on free speech, governance, and digital rights. Given Meta's significant presence in the region, this study aims to examine how AI biases in content moderation influence public discourse and affect democratic principles. By analyzing these issues, the paper seeks to uncover the complex relationship between technology and governance that shapes AI moderation practices in the Global South. Ultimately, it calls for a more nuanced approach to content moderation that respects regional differences and upholds core democratic values like freedom of speech.

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

2. Content moderation: a core function of social media platforms

Content moderation usually refers to the practice of determining which categories of content are allowed and prohibited on their platforms (Boberg et al., 2018; George & Scerri, 2007; Sander, 2019). It involves making decisions about which types of content are permissible, and which are not. This serves as a form of *content gatekeeping*, where the platform sets rules and guidelines to filter out content that violates its policies (Gillespie, 2018). For instance, platforms prohibit terrorist content, hate speech, defamation, harassment, or graphic violence. Howard points out, the purpose of content moderation is defensive; it seeks to mitigate the risks posed by ongoing threats and protect potential victims from (wrongful) harm, thus social media companies have a natural obligation to engage in some level of content moderation to safeguard those who are harmed or endangered by wrongful speech (Howard, 2023). Content moderation decisions of social media platforms are usually based on platform rules or community standards, architectural designs, algorithms, or human reviews that vary depending on the size, resources, purpose, and culture of the platform (Gillespie et al., 2020; Gongane, Munot & Anuse, 2022; Reuber & Fischer, 2022).

It has been argued that the evolution of platform liability regimes has enhanced the effectiveness of content moderation by holding platforms accountable for hate speech shared by users (Armijo, 2021; Hartmann, 2020; Langvardt, 2017; Sander, 2019). A prominent example is the EU's recently enacted Digital Services Act (DSA), which requires major platforms to implement systems to receive notifications of alleged illegal hate speech (European Commission, 2024). These platforms are obligated to review most valid notifications within 24 hours and, where appropriate, remove or disable access to such content (European Commission, 2024). This legal framework aims to ensure swift and efficient moderation, fostering a safer online environment while upholding regulatory standards. However, this concern has been opposed and criticized by a few other scholars (Brown, 2020; Frosio & Geiger, 2023; Lefouili & Madio, 2022; Sander, 2019). While procedural rules for content moderation usually receive more lenient scrutiny than substantive ones, the DSA's extensive requirements create a burden on platforms, potentially conflicting with the recent U.S. state laws (Nunziato, 2023).

In this regard, the United States follows a free speech absolutism model with a corporate self-regulation approach (Clifford, 2021; Foot, 2014). The U.S. legal framework strongly prioritizes free speech protections, as enshrined in the First Amendment, with Section 230 of the Communications Decency Act – labeled as the "safe harbor" provision, shielding platforms from liability and granting them broad discretion in content moderation (Armijo, 2021; Chander, 2022; Citron & Wittes, 2018). The scope of the Section 230 has been interpreted broadly in favor of the private platforms by the courts. For example, In *Zeran v. AOL*, the Fourth Circuit ruled that AOL was not liable for defamatory third-party posts, affirming Section 230's shield against defamation claims (*Zeran v. America Online, Inc*, Zeran v, 1997). Similarly, in *Jones v. Dirty World*, the Sixth Circuit held that platforms could curate content without losing immunity, as long as they did not materially contribute to its illegality (*Jones v. Dirty World Entertainment Recordings LLC*, v, 2014). While this approach ensures a high degree of speech freedom, it also allows harmful content, including disinformation and extremist propaganda, to persist unchecked. The reliance on corporate self-regulation has resulted in inconsistent enforcement, often prioritizing business interests over user rights (Gleiss et al., 2023; Ranchordás, 2019).

Similarly, as an emerging country from the global south India introduced to the regulatory framework in 2021, following a hybrid regulatory model that combines constitutional protections for free speech with significant government intervention in online content oversight (Abhishek, 2023; Karanicolas, 2021). While the legal framework recognizes the right to freedom of expression, the Information Technology (IT) Rules (2021) grant the government broad discretionary powers to mandate content removal, raising concerns about political suppression and the lack of transparency in enforcement (Abhishek, 2023). Additionally, the requirement for traceability in end-to-end encrypted communications poses a serious threat to digital privacy, leading to conflicts

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

between regulatory objectives and fundamental human rights protections (Nojeim & Maheshwari, 2021). This model reflects a delicate balance between the regulatory measures and free speech but also highlights the risks of excessive state control over digital platforms. However, earlier, in *Shreya Singhal v. Union of India*, the Supreme Court of India struck down Section 66A of the IT Act for its vagueness and potential chilling effect on free speech (*Shreya Singhal v. Union of India*, Shreya Singhal v, 2015). This landmark ruling established a critical precedent, affirming that the government's interest in regulating unlawful content must be balanced against constitutional freedoms. The decision set the stage for heightened judicial scrutiny of online speech regulations, emphasizing that laws governing cyberspace must avoid broad or imprecise language that could suppress legitimate discourse. This standard will likely inform the application of the IT Rules, 2021.

As the government slowly moves towards regulating speech on social media through various frameworks, private platforms are operating in the content moderation more seriously . Platforms generally carry out their content moderation primarily in two ways: manual (human) and automated (Binns et al., 2017; Gorwa et al., 2020). Human moderation involves individuals (moderators) manually reviewing and assessing content. In contrast, algorithmic moderation utilizes algorithms and machine learning to analyze and filter content across SMPs automatically. While some harmful speech content can be effectively identified through automated systems, these enforcement mechanisms often lack sensitivity to context and provide minimal explanations for their decisions (Gorwa et al., 2020; Hamon et al., 2020; Kiritchenko et al., 2021). Consequently, human review becomes essential when contextual cues are crucial for effective enforcement (San Martín, 2023). Although AI is an emerging model for content moderation due to its ability to process large volumes of content quickly and consistently, it faces numerous challenges. These challenges are particularly evident as regulatory approaches increasingly rely on AI-based systems, which is discussed in the next part of the paper. Therefore, this model is supplemented by human moderators who play a crucial role in ensuring more attentive oversight. For example, in the January 2024 Senate hearing on child safety concerns, the CEOs of leading SMPs – including Meta (formerly Facebook), X, TikTok, Snapchat, and Discord – recognized the necessity of integrating human moderators to tackle the limitations posed by algorithmic content moderation (Paul, 2024).

3. Meta's algorithmic content moderation

3.1 Algorithmic content moderation

Algorithmic content moderation refers to the use of automated systems, such as machine learning models, keyword filters, and pattern recognition tools, to monitor, filter, and regulate user-generated content on online platforms (Grimmelmann, 2015; Seering et al., 2019). Contemporary studies build upon the foundational standards of algorithmic content moderation established by early scholars, taking diverse approaches and directions. For example, Gorwa et al. adopt a narrower focus on commercial content moderation, which refers to automated systems that classify user-generated content through matching or prediction, resulting in enforcement actions such as removal, geoblocking, or account takedown (Gorwa et al., 2020). Though platforms claim that content moderation is supported by human moderators, in practice, it heavily relies on AI due to the sheer volume of content shared daily across the platforms (Gillespie, 2018; Sander, 2019).

AI content moderation creates massive discriminatory approaches in moderating content, especially from the Global South (Gorwa et al., 2020; Marsoof et al., 2023; Shahid & Vashistha, 2023). An important criticism has been that the content moderated and/or regulated through these platforms is determined by a relatively homogenous group of Silicon Valley *elites* (Greene, 2019; Medzini, 2022). Even though the Global South is rapidly becoming the largest emerging market for social media (Ghai et al., 2022), few other digital law scholars have argued that the practice of content moderation remains largely shaped by the cultural norms of the United States, given that most tech platforms

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

are headquartered there (Kaye et al., 2022; Kwet, 2019). Griffin, Patricia, and Cornell find that this imbalance in content moderation procedures has led to the proliferation of inequality, disparities, and occasional arbitrary practices in governing social media content within the global south (Griffin, 2023; Waldron & Ann, 2023).

Why does this difference occur? The disparities in content moderation arise from various factors, including cultural prejudices, economic inequalities, language barriers, and power dynamics (Rowe, 2022; Takhshid, 2021). Since the training of AI and large anguage models (LLMs) is central to algorithmic moderation, Gabriel Nicholas and Aliya Bhatia, in their work for the Center for Democracy & Technology (CDT), highlight the disparities in language resources (Nicholas & Bhatia, 2023). The study argues that, high-resource languages like English benefit from extensive digitized content, well-funded computational linguistics research, and advanced machine translation tools. In contrast, low-resource languages face significant challenges, including limited structured data, poor translation tools, and inadequate institutional support for natural language processing (NLP). Tech companies often rely on machine translation rather than investing in native language resources, leading to errors, cultural inaccuracies, and biases. As a result, content moderation on private platforms disproportionately impacts minority and local languages. For example, platforms like Facebook and Twitter struggle with moderating content in languages, such as Burmese (Myanmar), Amharic (Ethiopia), and Sinhala/Tamil (Sri Lanka), allowing misinformation and hate speech to go unchecked (Nicholas & Bhatia, 2023).

De Gregorio and Stremlau also point out this central issue from a different view point arguing,

> The imbalance of economic power between the Global South and social media companies – their valuations can be many times the GDP of poorer countries – means the tech titans pay little heed to the concerns of hate speech or disinformation campaigns in such nations, not least because they are marginal markets (De Gregorio & Stremlau, 2023).

Recent scholarly discussions from the Global South have increasingly addressed biases in social media content moderation, highlighting the need for context-aware frameworks. Scholars like Gondwe and Nemer have explored mundane technology, as exemplified by Zambian youth, as a crucial lens for examining these biases (Gondwe, 2024; Nemer, 2024). This perspective emphasizes how marginalized groups leverage everyday technological skills to navigate and resist government censorship, surveillance, and algorithmic control. By centering grassroots strategies, it exposes how these communities deconstruct dominant technological paradigms and repurpose them to suit their specific sociopolitical contexts, challenging elite control over digital spaces.

With this background, the following section examines various case studies related to Meta's AI content moderation and its challenges in the Global South. The choice to focus on Meta is influenced by the fact that Facebook is one of the most widely used sSMPs in this region (Poushter, 2024).

3.2. Meta's AI moderation and intrigue challenges

Lyons' study highlights that Facebook endangered users in the Global South by utilizing them as test subjects for underdeveloped content moderation policies prior to deploying those policies during the tumultuous U.S. election (Lyons, 2020). This practice exemplifies a broader trend among major SMPs, which frequently lack the adequate moderation tools and oversight mechanisms necessary to effectively manage harmful content in widely spoken non-Western languages. The absence of such tools exacerbates risks of understanding the context of the content shared by the users in these regions, as these platforms fail to provide the same level of protection that users in Western contexts receive.

Nourooz Pour criticizes Meta's content moderation practices and their impact on human rights standards, highlighting the challenges and limitations of its current moderation framework, pointing out,

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

> First, Meta's community standards and internal guidelines suffer from ambiguous language. Despite aiming for universal applicability, this vagueness hinders the provision of the contextual clarity essential for consistent and fair enforcement. Second, notwithstanding their much-vaunted 'context-sensitivity', these algorithms are limited by platform's categorical approach to assessing harm risk. Tey sort content, users, or actions into Meta's predefined categories and then apply rules based on these classifications. Tis setup, while capturing the literal meanings of words, often fails to grasp the nuanced contexts, local idioms, sarcasm, and cultural connotations inherent in the content (Nourooz Pour, 2024).

For example, the ambiguity of Meta's content moderation policies is evident in how these policies shape online narratives, particularly through the lens of global security concerns (Meta, 2025a). Security labels such as *Violence and Incitement, Dangerous Individuals and Organizations*, and *Coordinating Harm and Promoting Crime* disproportionately impact freedom of speech, especially for content generated in non-English languages. This challenge is further exacerbated by the fact that the multilingual language models used by Meta for content moderation perform significantly better in high-resource languages like English, Spanish, and Chinese, compared to medium and low-resource languagesc (Nicholas & Bhatia, 2023). This performance gap highlights the disparity in available training data and investment across different languages.

Access Now also has pointed out that Meta has consistently underinvested in content moderation for non-English speaking markets, which has contributed to the proliferation of hate speech, disinformation, and incitement to violence (Okkonen, 2024). Also, despite making public commitments, Meta has failed to provide adequate transparency regarding its content moderation practices and human rights impact assessments. This lack of transparency raises significant concerns about the effectiveness and accountability of the platform's moderation efforts, particularly in regions where the consequences of online harm are often amplified.

Furthermore, despite Meta's claim that it hires reviewers with specific language proficiency and cultural competency for different regions, the exact distribution of moderators by language remains undisclosed in the search results (Meta, 2022). This lack of transparency raises significant questions about the content moderation process and the effectiveness of these practices across diverse linguistic and cultural contexts.

Griffin argues that discrimination is likely to persist due to biases embedded in AI systems, which often make discriminatory decisions based on skewed data or flawed algorithms used in automated social media and application processes (Griffin, 2023). Leung also pinpoints this issue,

> Algorithms work better in languages that are frequently used, and less well in minority languages. More - over, Meta allocates unequal resources and prioritises attention to cases based on the urgency to control bad press (Leung, 2022).

Due to these complexities, algorithmic decisions often reflect inherent biases, producing inconsistent and frequently discriminatory outcomes across the global demographics and audience. Moreover, existing studies suggest that platforms intentionally leverage their moderation tools and algorithms to marginalize underrepresented voices, especially when these voices engage in discussions on global or international issues (Abokhodair et al., 2024; Vaccaro et al., 2021).

Additionally, due to the lack of contextual awareness, in the global applications, as the social media spaces become avenues for the modern-day expression of free speech, the content moderation decisions – often ends up with the wrong or false accusations – invade the free speech shared across these spaces. The main challenge in regulating 'online speech' stems from the complex, personal, and context-sensitive nature of language (Kadri & Klonick, 2019; Sander, 2019). Even commonly recognized slurs can be reclaimed by certain communities, changing their meaning and intention. When systems lack context awareness, they can make oversimplified judgments, flagging content for moderation by individuals who may not fully understand whether the speaker belongs to the group

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

being referenced by the supposed hate speech (Gorwa et al., 2020). For example, COFEM alleges that tech companies, including Meta, play a significant role in perpetuating a culture of digital violence, and their failure to address online gender-based violence (OGBV) not only facilitates this issue but also creates a cycle of repression that stifles free expression and undermines the ongoing struggle for gender equality (COFEM, 2024).

While online speech is increasingly suppressed by AI-driven moderation processes, these systems often lack the capacity to understand the nuances of content and context fully. As a result, the type of "harmful" content to be removed – e.g., violent rhetoric, misinformation, hate speech – is left to the SMPs' assessment. The necessity of removing harmful content from platforms is underscored by the potential repercussions on the social media ecosystem when such standards are not upheld. Chew and Tandoc Jr., in their study on Facebook Live's handling of the Christchurch shooting footage, illustrate this issue: The video remained accessible for a time, leading to public criticism pointing out that removing violent content is crucial to prevent desensitization to violence and mitigate the risk of psychological harm, highlighting the broader need for prompt content moderation to maintain a responsible digital environment (Chew & Tandoc, 2020).

In addition, AI-driven decisions are not always accurate, as these systems often create significant disparities by failing to grasp the context in which content is generated. When moderation is perceived as unjust, users may feel that the platform misunderstood the legitimacy of their posts and failed to protect their right to free expression. Though platform policies comprehensively address many challenges, the use of automated systems for content identification is cited as a potential source of errors. These systems may generate false positives or fail to recognize nuanced contextual differences that determine whether a piece of content actually violates platform guidelines (Ajder & Glick, 2021). Indeed, without any form of content moderation, users would likely find it difficult to participate in, or even desire to engage with, these online communities, given the propensity for some individuals to inundate digital public spaces with violent, extremist, vulgar, misleading, and spam-laden content (Thacker, 2023).

Pour points out, Meta regularly consults with various stakeholders, including civil society organizations, activist groups, and experts in fields like digital rights, free speech, and human rights, to identify potential concerns. While this practice is admirable from a human rights standpoint, the interpretations from these stakeholders often lack a clear link to national laws, international standards, or even Meta's own guiding principles (Nourooz Pour, 2024). O'Kane also makes this point; Meta's speech governance does not operate within an international human rights law (IHRL) framework. The company has moved away from its initial "'post-as-trumps" stance, which was rooted in a First Amendment-style "classic libertarian ethos" that emphasized minimal interference with user content and favored unrestricted individual expression (O'Kane, 2021).

As an evident, in a study, Human Rights Watch has identified six recurrent patterns of undue censorship performed by Meta, each occurring at least 100 times. These include: (1) the removal of posts, stories, and comments; (2) the suspension or permanent disabling of user accounts; (3) temporary restrictions on user engagement with content, such as liking, commenting, sharing, or reposting, which can last from 24 hours to 3 months; (4) limitations on the ability to follow or tag other accounts; (5) restrictions on the use of specific features, including Instagram/Facebook Live, monetization options, and the recommendation of accounts to non-followers; and (6) "shadow banning," characterized by a significant reduction in the visibility of an individual's posts, stories, or account, without prior notification, which can result from decreased distribution or reach of content and the disabling of search visibility for accounts (Human Rights Watch, 2023).

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

3.3. Meta's AI bias through case studies

In drawing the thin layer to keep content and remove content in the online spaces, as the core duty of content moderation process, SMPs, particularly Facebook, have been criticized for implementing discriminatory and unfair content moderation measures targeting the Arab world (Alimardani & Elswah, 2021; Díaz & Hecht-Felella, 2021; Gorwa et al., 2020). Facebook's organizational structure in this region reveals systemic issues that align with broader orientalist tropes (Alimardani & Elswah, 2021). While countries like Israel and most European nations have dedicated public policy heads, the Middle East and North Africa (MENA) region – despite its vast linguistic, cultural, and religious diversity – is managed under a single, generalized system. Although Facebook operates a general MENA office in Dubai, it has a specific office in Israel with its own public policy director, Jordana Cutler, who previously served as an adviser to former Israeli Prime Minister Benjamin Netanyahu and worked with the Likud party. In contrast, no equivalent roles exist for Palestinians or other Arab countries, exposing significant disparities in representation and attention (Alimardani & Elswah, 2021).

For instance, in May 2021, Facebook characterized the censorship of *pro-Palestine* content as mere "technical errors," a statement that believes deeper, more pervasive issues of systemic design discrimination (Fatafta, 2024). This response reflects a troubling trend, amplified by insufficient resources and discriminatory policies. Such practices have fostered situational crises within the Arab world, specifically as pro-Palestine voices confront increasing digital orientalism alongside tangible forms of repression. This environment not only curtails free expression but also reinforces a cycle of marginalization for advocates of Palestinian rights (Alimardani & Elswah, 2021). This shadow banning of content highlighted by Journalist Azmat Khan, stating,

> After posting an Instagram story about the war in Gaza yesterday, my account was shadowbanned. Many colleagues and journalists friends have reported the same. It's an extraordinary threat to the flow of information and credible journalism about an unprecedented war (Business & Human Rights Resource Centre, 2023).

Research by Access Now highlights Meta's uneven approach to election-related content moderation, with a disproportionate focus on the United States while neglecting non-Western, non-English-speaking markets (Okkonen, 2024). Despite media reports suggesting that Meta is implementing ad-related safeguards for the U.S. elections, the company has not disclosed adequate measures to protect voters elsewhere, leaving them vulnerable to hate speech, disinformation, and incitement to violence. In India, Meta has refused to publish the full human rights impact assessment of its preparedness for the 2024 elections, raising concerns about transparency and accountability. Similarly, a Global Witness investigation in Brazil exposed Meta's failure to prevent the spread of election-related disinformation on Facebook. Further underscoring these issues, the European Commission has initiated formal proceedings against Meta for potential violations of the Digital Services Act, highlighting ongoing regulatory compliance concerns.

For instance, Gizele Martins, a grassroots activist and human rights advocate from Brazil, sheds light on the challenges faced in the region:

> "The absence of basic human rights plus racism and social inequality makes us have to decide who gets to receive food amidst widespread hunger. I want the world to know that mutual support and solidarity are what is going to save us from any crisis while we still don't have the same right to life as the wealthy do (Fakomogbon, 2022).

This quote underscores the pressing struggles faced by activists in the Global South, highlighting the intersection of inequality, human rights, and social support in times of crisis.

Another example is Meta's engagement-based algorithms on Facebook have contributed to an *anti-Rohingya* echo chamber in Myanmar, amplifying inflammatory and hate-filled content for profit by

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

keeping users engaged. This environment allowed actors, including the Myanmar military and radical nationalist groups, to spread disinformation and incite violence against the Rohingya minority. Senior military figures, such as General Min Aung Hlaing, further escalated this hatred, denying the existence of the Rohingya and later seizing power. A study conducted by Amnesty International identified attempts to counter hate speech, like Meta's "Panzagar" initiative, backfired, with Facebook algorithms inadvertently promoting hateful content (Amnesty International, 2022).

In response, Facebook admitted its role in amplifying hate speech during the crisis and has since taken steps to address the issue (BBC, 2018). This includes efforts to increase content moderation, such as hiring more Burmese-speaking moderators to better manage content in the region and mitigate the harmful effects of online hate.

In a research study conducted by the Amnesty International around this incident, it was revealed that while the Rohingya faced systemic discrimination by Myanmar's authorities for many decades before 2012, they had coexisted relatively harmoniously with other ethnic groups in Rakhine state (Amnesty International, 2022). However, this changed with the increasing use of social media platforms, particularly Facebook, in the country. Mohamed Ayas, a Rohingya school teacher, reflected on this shift in a statement to Amnesty International:

> We used to live together peacefully alongside the other ethnic groups in Myanmar. Their intentions were good to the Rohingya, but the government was against us. The public used to follow their religious leaders, so when the religious leaders and government started spreading hate speech on Facebook, the minds of the people changed (Amnesty International, 2022).

Meta has also faced criticism for exacerbating hate speech related to religious tensions and violence in Bangladesh. This issue goes beyond the Hindu–Muslim conflict; it also encompasses the treatment of religious minorities, such as Buddhists and others, who have suffered violent incidents as a result of the difficulties in identifying content in the Bangla language (Hossain, 2023).

Recently, in 2025, the contemporary impact of Meta's policy changes on Africa, particularly in the context of content moderation, raises significant concerns. As noted by CIPESA (Collaboration on International ICT Policy for East and Southern Africa),

> Meta's decision is particularly concerning for Africa, which is unique in terms of linguistic and cultural diversity, limited digital and media information literacy, coupled with the growing challenges of hate speech and election-related disinformation, lack of context-specific content moderation policies, and inadequate investment in local fact-checking initiatives (CIPESA, 2025).

This highlights the challenges posed by Meta's policies, which are ill-suited to the region's complexities. A case study from Kenya further illustrates the impact of these changes: 'Meta's decision to abandon fact-checking raises critical concerns for Africa, coming after the tech giant's January 2023 decision to sever ties with their East African content moderation contractor, Sama, based out of Nairobi, Kenya, that was responsible for content moderation in the region.' This move has exacerbated the challenges of content moderation in the Global South, particularly in light of the region's unique needs for culturally sensitive and context-specific approaches.

3.4. Failed promises of Meta's solutions

In response to these content moderation challenges, Meta introduced the Oversight Board, an independent body that examines the cases of content moderation decisions across Facebook and Instagram platforms. However, the Oversight Board's system lacks the comprehensiveness needed to

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

establish an effective reconciliation mechanism for the complex issues arising from content moderation practices. Leung J criticizes Meta for facing difficult cases. These cases involve limited context, such as offline interactions or exchanges on other platforms. They also include value conflicts, like balancing diverse voices with user safety. Additionally, there are complex situational factors. However, Leung points out that even when relevant context is available, Meta often excludes it from its content analysis (Leung, 2022). The practices of contextualization, literalization, and monomodal orientation – where focus is placed on a single method or perspective – serve as interpretive shortcuts designed to enhance efficiency and scalability. However, these approaches often overlook or suppress critical information that could provide clearer meaning. As a result, uncertainties that could be easily resolved are left unaddressed.

Further, For example, in the case of *Protest in India Against France*, a majority of the Oversight Board did not find Facebook's contextual justification for possible violence in this specific instance to be compelling (Oversight Board, 2022). In contrast, a minority of the panel advocated for deferring to Facebook's assessment that the post posed an unacceptable risk of inciting violence, partly because the company had relied on a third-party partner assessment and consulted regional and linguistic experts (Helfer & Land, 2022).

In 2023, Meta's Oversight Board reported that 38% of content moderation cases originated from the United States and Canada, while 26% came from Europe—accounting for a total of 64% of all cases. In contrast, only 5% of cases were reported from Central and South Asia, despite India having the largest number of Facebook and Instagram users globally (Oversight Board, 2023). This stark disparity highlights the disproportionate focus on the Western world compared to the Global South, raising concerns about the unequal allocation of content moderation resources and oversight.

Paul M. Barrett pointed out that The Oversight Board's direct impact on Meta's moderation practices appears minimal if measured solely by its binding rulings. To date, the Board has issued only 53 rulings—representing a mere 0.0018% of the nearly 3 million moderation decisions that users unsuccessfully appealed to Meta and then escalated to the Board. However, the Board has consistently challenged Meta's decisions, overturning nearly 80% of cases, and its public opinions frequently expose inconsistent policies and ad hoc enforcement practices (Barret, 2023).

This lack of tie with the Meta and Meta's Oversight Board has been criticized by O'Kane who points out that,

> Meta also has a propensity to mistranslate the Board's recommendations and overstate the extent to which its responses in fact implement these recommendations. Meta has attempted to restrict recommendations to a specific case, even where the Board clearly intended the recommendation to be a broader policy recommendation, for instance regarding notification of enforcement actions (O'Kane, 2021)

These case studies and scholarly analyses underscore the persistent issues with Meta's AI-driven content moderation policies, their application, and the limitations of the Oversight Board, highlighting an urgent need for reform in policy and content regulation. These policies infringe on fundamental user rights – particularly freedom of speech and expression – while undermining democratic values and the public good, specifically within the Global South.

In response, various actors – including nongovernmental organizations and public authorities – have sought to tackle the harms caused by the spread of hate speech and disinformation online (de Cock Buning, 2018; Gagliardone et al., 2015; Griffin, 2022). These efforts aim to counter the negative effects of inadequate content moderation and foster a safer digital environment (De Gregorio & Stremlau, 2023). These methods involve both private and public authorities taking on various roles in social media content regulation, particularly the role of governments and private platforms. This approach sparks debate over public versus private governance in this domain, a topic explored in the upcoming section.

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

4. Current debates between private and public actors

Despite extensive legislative and policy efforts worldwide to address this persistent issue, it continues to prevail globally, with particularly severe impacts in the Global South. This enduring challenge arises from factors on both the platforms' and governments' sides, as they navigate the complex intersection of private and public speech governance on social media platforms.

From the SMPs' perspective, AI-based content moderation stems primarily from internal regulations and relies heavily on *voluntary* commitments. It remains challenging to discern the true motivations behind these frameworks – whether they are genuinely intended to uphold responsible practices or merely serve as superficial measures to *appease* external scrutiny. For example, Zahra highlighted the Meta's Oversight Board as a CSR Tactic, announced in 2018 that the Board was established to address complex issues of online freedom of expression, such as decisions on content removal or retention, Meta funded the Board through a trust and set it up as a Delaware limited liability corporation with both corporate and individual trustees (Takhshid, 2021).

Due to the voluntary nature of these initiatives, the processes, standards, and mechanisms employed by tech giants like Meta are neither impactful nor effective in addressing the core issues they claim to tackle. Rather than offering genuine solutions, these platforms often engage in superficial measures designed to create the appearance of action for users and the international community. In reality, such efforts do not address the root causes of the problems; instead, they resemble awareness campaigns rather than providing the necessary "treatment" for the underlying issues.

Second, a common criticism leveled at platforms is their tendency to approach operations primarily from a corporate standpoint, ultimately prioritizing commercial interests. As they are fundamentally profit-driven entities, this corporate-centric approach often shapes their policies and actions.

Helfer and Land argue that the free speech aspect of platforms are influenced by a range of factors, predominantly by the commercial objectives,

> Platforms also have different interests than governments: they pursue commercial objectives rather than societal welfare, and their own speech interests also play a role. It remains unclear how these interests should be accommodated within the human rights framework governing free expression (Helfer & Land, 2022).

Adding to this, Takhshid also underlined that the harmful consequences of social media platforms should not be overlooked in favor of protecting corporate interests, especially when these companies promote humanitarian ideals but fail to act responsibly when it matters most – citing the example of Meta posts having contributed to the violence in Ethiopia, illustrating how falsehoods spread on social media can have devastating real-world consequences without any accountability in 2020 (Takhshid, 2021).

Unfortunately, the business and profit models of these platforms are largely driven by content reach, which fuels multiple revenue streams, such as advertising revenue. This dependency on content engagement often makes platforms hesitant to remove content, even when faced with significant criticism or conflicts with existing policies and regulations.

For instance, Balkin criticizes the business model of these major platforms,

> The problem with the current business models for social media companies such as Facebook, Twitter, and YouTube is that they give companies perverse incentives to manipulate end users—or to allow third parties to manipulate end users—if this might increase advertising revenues, profits, or both (Balkin, 2018).

Currently, most of the major platforms rely on algorithms for content moderation. These algorithms are designed to adapt based on user behavior and preferences, aligning with the objective of maximizing user profiles. By prioritizing content that reflects users' interests and engagement patterns,

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

they aim to increase interaction. However, in select instances, profit-driven social media companies strategically ban users or restrict content as part of a broader economic calculation. The underlying rationale is to enhance the perceived value for marginal users. Within an advertisement-driven model, platforms are inclined to remove content only when doing so fosters greater user engagement, resulting in increased time spent interacting with content and advertisements. Ultimately, these actions are intended to optimize revenue within the framework of the platform's business model (Jiménez-Durán, 2023).

Third, from the perspective of these platforms, as private corporations, their corporate policies play a pivotal role in shaping the content allowed or restricted on their sites. In structuring their platforms, these companies tailor their moderation practices to align with the environments they wish to create, enforcing their policies with varying degrees of strictness depending on specific contexts. Such approaches often have a direct impact on users' freedom of expression. For instance, during the Black Lives Matter protests, Facebook implemented several restrictive measures in its content moderation practices, citing violations of privacy, community standards, or platform policies, which critics argue disproportionately affected marginalized voices.

For example, in January 2025, Meta introduced significant changes to its content moderation policies, sparking concerns about human rights and online safety (Meta, 2025b). The key shifts include replacing fact-checkers with a community notes system, simplifying content policies, raising the threshold for content removal, reintroducing political content, relocating Trust and Safety teams to Texas, and collaborating with President Trump to resist regulation. While framed as promoting free expression, these changes risk mainstreaming hate speech, undermining fact-checking, and increasing misinformation, particularly during elections. Notably, the revised "Hateful Conduct Community Standard" now permits exclusionary and derogatory language based on gender, sexual orientation, and national origin, contradicting Meta's human rights commitments. Replacing expert fact-checkers with community-driven moderation raises concerns about misinformation control, as such systems may lack the expertise and consistency of professional oversight. Additionally, the relaxation of content removal standards and the return of political content without safeguards heighten the risk of harmful narratives spreading unchecked. Stefania Di Stefano has criticized these shifts, arguing that Meta's collaboration with President Trump to resist regulation raises serious concerns about its commitment to accountability and responsible governance (Di Stefano, 2025). Given Meta's past failures in content moderation, such as in Myanmar, these policy changes reflect a troubling departure from its previous efforts to align with human rights principles.

ARTICLE 19 also has strongly criticized this Meta's content moderation changes, highlighting their politically motivated timing (ARTICLE19, 2025). The announcement, made just before President-elect Trump's inauguration, appears to prioritize appeasing conservative political interests over a genuine commitment to free expression. ARTICLE 19's call for Meta to uphold human rights over political posturing aligns with broader concerns from experts and advocacy groups. The company's decision to collaborate with Trump to "fight censorship" while simultaneously loosening content moderation policies raises serious questions about platform accountability and user safety. These developments underscore the persistent challenge of balancing free speech with the need to mitigate online harms and emphasize the importance of transparent, accountable content moderation practices that respect human rights for all users.

Fourth, the lack of integration between platforms and their resolution mechanisms has been identified as a key issue, as these solutions often appear hastily implemented and insufficient to address the ongoing crises and disruptions within the platform ecosystem. For instance, the significant disconnect between Meta's policies and the Meta Oversight Board, as discussed in the previous section, highlights this issue. Policy scholars argue that for the Oversight Board to function effectively as a constitutionalising instrument within Meta's virtual communities – namely Instagram and Facebook – it must expand its engagement with content moderation policies (Filatova-Bilous, 2023; Quintais et al., 2023). This expansion should adopt a more inclusive approach, incorporating the perspectives

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

of users and civil society organizations, specifically from underrepresented regions. By doing so, the Board can better address the diverse challenges users face worldwide, fostering a more equitable and representative framework for content governance across these platforms (Da Conceição, 2023).

Fifth, while many Big Tech companies have advocated for the ethical use of AI, a significant lack of accountability and oversight remains in numerous countries within the Global South. Despite this, the economic growth of these countries will largely depend on the data that social media companies have amassed over the years. Holding such a vital resource gives these companies considerable leverage, not only over other private enterprises but also over governments that may not fully grasp the critical importance of the data collected by social media platforms (Takhshid, 2021).

Similarly, from the governments' perspective, policy scholars suggest that addressing the inequalities and disparities in content moderation requires more than a singular approach. A hybrid strategy that integrates bottom-up, top-down, and international methods is essential to achieve meaningful progress (De Gregorio & Stremlau, 2023). From a top-down perspective, however, many governments in the Global South face significant challenges in effectively addressing content moderation. Challenges in regulating platforms from the governmental perspective persist as a significant struggle, exacerbated by a lack of capacity and appropriate tools among many governments to address these issues effectively. In response, a common initial reaction has been to criminalize the dissemination of online hate and disinformation among users and social media platforms. Some governments increasingly view the proliferation of such content as a justification for censoring speech and even shutting down the Internet (Howard et al., 2011; Marchant & Stremlau, 2020). For example, India has been recorded with the most number of the internet shutdowns, with 771 blackouts from 2016 to 2023 (Bhattacharya, 2024). These shutdowns, often aimed at controlling public dissent and managing civil unrest, have been implemented to suppress protests related to the Citizenship Amendment Act, curb the farmers' protests, and even prevent exam cheating (Ruijgrok, 2022; Shah, 2021).

This approach is often perceived as an immediate and effective remedy to mitigate escalating violence, particularly in light of companies' discretionary responses to content moderation. However, there is limited evidence to support the effectiveness of these practices in combating the misinformation and hate they aim to address. Yet, shutdowns have been implemented as a means to restrict online speech (De Gregorio & Stremlau, 2023).

Legislative initiatives from the region of the Global South indicate a rising interest in addressing the challenges of online content moderation. There is a significant push for regulatory bills that advocate for strong state intervention, often at the expense of due process. This trend highlights the critical necessity for legislative, administrative, and judicial bodies to uphold minimum standards for protecting free speech (Lanza & Jackson, 2021). It also highlights the complications arising from the platforms' internal regulatory mechanisms and AI-driven content moderation, along with their subsequent repercussions.

For example, in India, in addition to the IT Rules discussed earlier in this paper, several key internet regulations were introduced in 2023 to modernize the digital governance framework and expand government control over online content and services. India, being one of the largest markets for SMPs, has seen significant regulatory changes. The Digital India Act 2023 aims to replace the outdated IT Act 2000, addressing contemporary challenges such as user harm, online security, and misinformation. The Telecommunications Act 2023 has overhauled the telecommunications regulatory framework, granting the government broad powers over the sector. Additionally, the IT Rules, with subsequent amendments, have intensified government oversight by mandating grievance redressal mechanisms and the identification of message originators. A fact-checking unit was established in 2023 to flag 'fake, false, or misleading' content, requiring platforms to remove such content within 72 hours.

Yet, there is massive reluctance to incorporate the legislative framework for the content moderation or governance in the fragile democratic States of the Global South – which also stretches the problem ongoing. This struggle to seek the support of the governments of the Global South States pointed by Zahra,

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

> There may be those who are rightly hesitant about the involvement of governments. This distrust is exacerbated by the involvement of governments with different constitutions and power dynamics, many of which are unfortunately autocratic. Even in the United States, with a two-hundred-year-old Constitution, many commentators and scholars are reluctant to fully delegate power to regulate social media companies to the government concerning a range of issues (Takhshid, 2021).

Gorwa et al. also highlight the complexities of regulatory governance, noting that these challenges are pronounced even in Western jurisdictions. As government pressure on technology companies escalates, both companies and legislators are increasingly pursuing technical solutions to address intricate content governance issues. Recent regulations, such as Germany's NetzDG and the EU Code of Conduct on Hate Speech, mandate stringent timelines for content removal, compelling platforms to rely heavily on automated systems for proactive detection. This reliance raises significant concerns about the adequacy and fairness of these methods (Gorwa et al., 2020).

However, due to these complications in setting-up a hybrid model, the complete involvement or replacement of private actors in speech regulation across social media would pose a risk of excessive censorship of online speech through stringent suppression regulations.

Pamela rightly notes,

> It is important to consider that throughout history, limiting freedom of expression is usually one of the first measures taken by autocratic or weak democratic governments, particularly the freedom of expression of those who criticize or oppose them. Although political speech is one of the most protected expressions under IHRL, it is the first to be censored by authoritarian regimes (San Martín, 2023).

In this case, the overall objective of transitioning to a system that protects online freedom of speech could become compromised. While the intention is to address how platforms' content moderation policies infringe upon free speech, the result may instead be an intensified suppression of expression that threatens the very foundations of free speech itself.

5. Concluding observations

Despite substantial legislative and policy efforts, AI content moderation challenge remains a challenging global issue, especially impacting the Global South, which is explored in this paper giving special reference to Meta. SMPs, primarily motivated by profit, often implement AI-driven content moderation measures that appear more superficial than substantive, focused on image management rather than addressing the root causes of harmful content. The voluntary and self-regulated nature of these measures frequently limits their effectiveness, particularly in contexts where economic pressures and advertising revenue play dominant roles. This situation is exacerbated by government approaches in the Global South, where limited resources, lack of technical capacity, and occasional authoritarian tendencies lead to heavy-handed methods such as internet shutdowns and restrictive laws that may stifle legitimate speech.

Ultimately, the current approaches by both platforms and governments create a precarious balance that risks undermining free speech, often prioritizing corporate and state interests over the safeguarding of online expression rights. Achieving genuine change will require integrated, hybrid approaches that incorporate the perspectives of both users and civil society, particularly in underrepresented regions, while carefully balancing the complex dynamics of commercial and regulatory interests.

Acknowledgements. The author wishes to thank the organizers and participants of the JUST-AI Jean Monnet Summer Colloquia (2024) for their valuable ideas and feedback, particularly Ljupcho Grozdanovski, Jérôme De Cooman, and Quentin Bebronne.

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

Funding. None declared.

Competing interests. None declared.

References

Abhishek, A. (2023). The state deputizing citizens to discipline digital news media: The case of the IT rules 2021 in India. *Digital Journalism*, *11*(10), 1769–1787. doi: 10.1080/21670811.2022.2134163

Abokhodair, N., Skop, Y., Rüller, S., Aal, K., & Elmimouni, H. (2024). Opaque algorithms, transparent biases: Automated content moderation during the Sheikh Jarrah Crisis. *First Monday*. Retrieved December 19, 2024, from https://firstmonday.org/ojs/index.php/fm/article/view/13620

Ajder, H., & Glick, J. (2021). Just joking! Deepfakes, satire and the politics of synthetic media. *WITNESS and MIT Open Documentary Lab. Retrieved April, 13*, 2022.

Alimardani, M., & Elswah, M. (2021). Digital orientalism:# SaveSheikhJarrah and Arabic content moderation. *Alimardani, Mahsa and Elswah, Mona. Digital Orientalism:# SaveSheikhJarrah and Arabic Content Moderation (August 5, 2021). In POMEPS Studies, 43*. Retrieved October 13, 2024, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3900520

Amnesty International. (2022, September 29). *Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations – New report*. Retrieved April 22, 2024, from https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Anderson, C. W. (2021). Fake news is not a virus: On platforms and their effects. *Communication Theory*, *31*(1), 42–61. doi: 10.1093/ct/qtaa008

Armijo, E. (2021). Reasonableness as censorship: Section 230 reform, content moderation, and the first amendment. *Florida Law Review*, *73*, 1199.

ARTICLE19. (2025, January 7). *Meta: Prioritise human rights, not politics*. ARTICLE 19. https://www.article19.org/resources/meta-prioritise-human-rights-not-politics/

Balkin, J. M. (2018). Fixing Social Media's Grand Bargain. *Hoover Working Group on National Security, Technology, and Law, Aegis Series Paper*, *1814*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3266942

Barret, P. M. (2023). *Meta's Oversight Board and the Need for a New Theory of Online Speech*. Default. https://www.lawfaremedia.org/article/meta-s-oversight-board-and-the-need-for-a-new-theory-of-online-speech

BBC. (2018, November 6). *Facebook admits it was used to "incite offline violence" in Myanmar*. https://www.bbc.com/news/world-asia-46105934

Bhattacharya, A. (2024, September 27). *India shuts down the internet far more than any other country*. Rest of World. https://restofworld.org/2024/india-internet-shutdown-record/

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In G. L. Ciampaglia, A. Mashhadi & T. Yasseri (Eds.), *Social Informatics* (vol. 10540, 405–415). Springer International Publishing. doi: 10.1007/978-3-319-67256-4_32

Boberg, S., Schatto-Eckrodt, T., Frischlich, L., & Quandt, T. (2018). The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication*, *6*(4), 58–69. doi: 10.17645/mac.v6i4.1493

Bradford, A. (2023). *Digital empires: The global battle to regulate technology*. Oxford University Press. https://books.google.com/books?hl=en&lr=&id=ibnQEAAAQBAJ&oi=fnd&pg=PP1&dq=Recent+events,+including+the+Christchurch+Mosque+shootings,+violence+against+the+Rohingya+population,+and+numerous+disinformation+campaigns+related+to+the+Brexit+referendum+and+the+2016+U.S.+elections,+as+well+as+Russia%E2%80%99s+invasion+of+Ukraine,+exemplify+the+pressing+need+for+regulatory+measures+addressing+speech+on+social+media.+&ots=b5Gy6Rk63N&sig=U9xdUMtxeuSWzSK58cKMA74HBGI

Brown, N. I. (2020). Regulatory goldilocks: Finding the just and right fit for content moderation on social platforms. *Texas. A&M Law Review*, *8*, 451.

Business & Human Rights Resource Centre. (2023, October 16). *Meta responds to allegations of Instagram shadow-ban for pro-Palestine content*. https://www.business-humanrights.org/en/latest-news/meta-responds-to-allegations-of-instagram-shadow-ban-for-pro-palestine-content/

Ceron, A. (2015). Internet, news, and political trust: The difference between social media and online media outlets. *Journal of Computer-Mediated Communication*, *20*(5), 487–503. doi: 10.1111/jcc4.12129

Chander, A. (2022). Section 230 and the International Law of Facebook. *Yale Journal of Law and Technology*, *24*, 393.

Chetty, N., & Alathur, S. (2018). Hate speech review in the context of online social networks. *Aggression and Violent Behavior*, *40*, 108–118. doi: 10.1016/j.avb.2018.03.003

Chew, M., & Tandoc, E. C. (2020). Lives and livestreaming: Negotiating social media boundaries in the Christchurch terror attack in New Zealand. In *Critical Incidents in Journalism* 178–189. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003019688-18/lives-livestreaming-matthew-chew-edson-tandoc

CIPESA. (2025, January 14). What Does Meta's About-Turn on Content Moderation Bode for Africa? *Collaboration on International ICT Policy for East and Southern Africa (CIPESA)*. https://cipesa.org/2025/01/what-does-metas-about-turn-on-content-moderation-bode-for-africa/

Citron, D. K., & Wittes, B. (2018). The problem isn't just backpage: Revising section 230 immunity. *Georgetown Law Technology Review*, *2*(2). doi: 10.2139/ssrn.3069747

Clifford, B. (2021). Moderating Extremism: The state of online terrorist content removal policy in the United States. *Program on Extremism*. *Washington DC*: *George Washington University*.

COFEM. (2024, May 20). *Silencing voices, erasing narratives: The battle against online gender based violence in the SWANA region*. https://cofemsocialchange.org/silencing-voices-erasing-narratives-the-battle-against-online-gender-based-violence-in-the-swana-region/

Da Conceição, L. H. M. (2023). A constitutional reflector? Assessing societal and digital constitutionalism in Meta's Oversight Board. *Global Constitutionalism*, 1–34. doi: 10.1017/S2045381723000394.

De Abreu Duarte, F. M. (2024). *The digital equilibrium: How governments, corporations, and individuals bargained the regulation of online speech in the European Union* [PhD Thesis, European University Institute]. https://cadmus.eui.eu/handle/1814/76743

de Cock Buning, M. (2018). *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. Publications Office of the European Union. https://cadmus.eui.eu/handle/1814/70297

De Gregorio, G., & Stremlau, N. (2023). Inequalities and content moderation. *Global Policy*, *14*(5), 870–879. doi: 10.1111/1758-5899.13243

Díaz, Á., & Hecht-Felella, L. (2021). Double standards in social media content moderation (1–23). *Brennan Center for Justice at New York University School of Law*.

Di Stefano, S. (2025, January 23). *Zuckerberg's "updated" recipe for Meta: "Prioritize speech" and neglect human rights*. OpenGlobalRights. https://www.openglobalrights.org/zuckerbergs-updated-recipe-for-meta-prioritize-speech-and-neglect-human-rights/

Elejalde, E., Ferres, L., & Herder, E. (2018). On the nature of real and perceived bias in the mainstream media. *PloS One*, *13*(3), e0193765.

Errington, K., Rahman, A., Barraclough, T., Barnes, C., Kukutai, T., Cormack, D., … **Judd, S.** (2020). *Shouting zeros and ones: Digital technology, ethics and policy in New Zealand* (vol 83). Bridget Williams Books. https://books.google.com/books?hl=en&lr=&id=azD3DwAAQBAJ&oi=fnd&pg=PR2&dq=Recent+events,+including+the+Christchurch+Mosque+shootings,+violence+against+the+Rohingya+population,+and+numerous+disinformation+campaigns+related+to+the+Brexit+referendum+and+the+2016+U.S.+elections,+as+well+as+Russia%E2%80%99s+invasion+of+Ukraine,+exemplify+the+pressing+need+for+regulatory+measures+addressing+speech+on+social+media.+&ots=NYxDhzoGLb&sig=eCAr7NyO1UmucnH4_XexNAObIqs

European Commission. (2024). *Questions and answers on the Digital Services Act**. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348

Fakomogbon, G. (2022, July 7). *27 Inspirational Quotes About Activism From Some of the World's Leading Activists*. Global Citizen. https://www.globalcitizen.org/en/content/inspirational-quotes-activism-activist-leaders/

Fatafta, M. (2024). *How Meta censors Palestinian voices*. https://www.accessnow.org/publication/how-meta-censors-palestinian-voices/

Filatova-Bilous, N. (2023). Content moderation in times of war: Testing state and self-regulation, contract and human rights law in search of optimal solutions. *International Journal of Law and Information Technology*, *31*(1), 46–74. doi: 10.1093/ijlit/eax041

Foot, K. (2014). The online emergence of pushback on social media in the United States: A historical discourse analysis. *International Journal of Communication*, *8*(0), Article 0.

Frosio, G., & Geiger, C. (2023). Taking fundamental rights seriously in the digital services act's platform liability regime. *European Law Journal*, *29*(1–2), 31–77. doi: 10.1111/eulj.12475

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing. https://books.google.com/books?hl=en&lr=&id=WAVgCgAAQBAJ&oi=fnd&pg=PA3&dq=These+frameworks+determine+which+content+is+permissible+and+which+is+prohibited,+thereby+influencing+the+dynamics+of+free+speech+and+expression+online&ots=Td5j6gGLOz&sig=OIkiL7ytRbTSdluoAN19pz-_sho

Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 913–922. doi: 10.1145/3178876.3186139

George, C. E., & Scerri, J. (2007). Web 2.0 and user-generated content: Legal challenges in the new frontier. *Journal of Information, Law and Technology*, *2*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1290715

Ghai, S., Magis-Weinberg, L., Stoilova, M., Livingstone, S., & Orben, A. (2022). Social media and adolescent well-being in the Global South. *Current Opinion in Psychology*, *46*, 101318.

Gillespie, T. (2017). Governance of and by platforms. *SAGE Handbook of Social Media*, 254–278.

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: CT Yale University Press. Retrieved January 6, 2024 https://books.google.com/books?hl=en&lr=&id=cOJgDwAAQBAJ&oi=fnd&pg=PA1&dq=Tarleton+Gillespie.,+Custodians+of+the+Internet:+Platforms,+Content+Moderation,+and+the+Hidden+Decisions+That+Shape+social+media+(New+Haven:+Yale+University+Press,+2018)+at+200.&ots=PiNNGSRPn5&sig=sAEAfuPcrrqrpEtinBwi0uExFoY

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

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.

Gleiss, A., Degen, K., & Pousttchi, K. (2023). Identifying the patterns: Towards a systematic approach to digital platform regulation. *Journal of Information Technology*, *38*(2), 180–201. doi: 10.1177/02683962221146803

Gondwe, G. (2024). Digital natives, digital activists in non-digital environments: How the youth in Zambia use mundane technology to circumvent government surveillance and censorship. *Technology in Society*, *79*, 102741. doi: 10.1016/j.techsoc.2024.102741

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. *Social Network Analysis and Mining*, *12*(1), 129. doi: 10.1007/s13278-022-00951-3

Gorwa, R. (2019). What is platform governance? *Information, Communication & Society*, *22*(6), 854–871. doi: 10.1080/1369118X.2019.1573914

Gorwa, R. (2024). *The politics of platform regulation: How governments shape online content moderation*. Oxford University Press. https://library.oapen.org/handle/20.500.12657/90834

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1), 205395171989794. doi: 10.1177/2053951719897945

Gosztonyi, G. (2023). *Censorship from Plato to Social Media: The complexity of social media's content regulation and moderation practices* (vol 61). Springer Nature. https://books.google.com/books?hl=en&lr=&id=uYjlEAAAQBAJ&oi=fnd&pg=PR8&dq=G%C3%A1bor+Gosztonyi,+Censorship+from+Plato+to+Social+Media:+The+Complexity+of+Social+Media%E2%80%99s+Content+Regulation+and+Moderation+Practices+(Vol+61)+(Springer+Nature,+2023)+at+45.&ots=RGj8G-ZdbL&sig=WjgvS4eQ14BPm_qAEz-DaKVQMz4

Greene, L. (2019). *Silicon states: The power and politics of big tech and what it means for our future*. Catapult. https://books.google.com/books?hl=en&lr=&id=bn1LEAAAQBAJ&oi=fnd&pg=PA3&dq=An+important+criticism+has+been+that+the+content+moderated+and/or+regulated+through+these+platforms+is+determined+by+a+relatively+homogenous+group+of+Silicon+Valley+elites.+&ots=JqIhkku8Sc&sig=Zlr-TRZ3LyBSZBRjScVvVw5cxxQ

Griffin, R. (2022). New school speech regulation as a regulatory strategy against hate speech on social media: The case of Germany's NetzDG. *Telecommunications Policy*, *46*(9), 102411. doi: 10.1016/j.telpol.2022.102411

Griffin, R. (2023). Rethinking rights in social media governance: Human rights, ideology and inequality. *European Law Open*, *2*(1), 30–56. doi: 10.1002/elo3.212

Grimmelmann, J. (2015). The virtues of moderation. *Yale Journal of Law and Technology*, *17*, 42–109.

Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, *207*, 2020. doi: 10.2760/124473

Hartmann, I. A. (2020). A new framework for online content moderation. *Computer Law & Security Review*, *36*, 105376. doi: 10.1016/j.clsr.2020.105376

Helfer, L. R., & Land, M. K. (2022). The meta oversight board's human rights future. *Cardozo Law Review*, *44*, 2233. doi: 10.2139/ssrn.3744510

Hossain, M. P. (2023). The implications of the Arabic case for Bangla content moderation on Facebook: Future considerations for combating 'Bangla' hate speech. *Journal of Cyber Policy*, *8*(3), 308–326. doi: 10.1080/23738871.2024.2337129

Howard, J. W. (2023). The ethics of social media: Why content moderation is a moral duty. *Journal of Practical Ethics*. https://discovery.ucl.ac.uk/id/eprint/10179623/

Howard, P. N., Agarwal, S. D., & Hussain, M. M. (2011). When do states disconnect their digital networks? Regime responses to the political uses of social media. *The Communication Review*, *14*(3), 216–232. doi: 10.1080/10714421.2011.597254

Human Rights Watch. (2023, December 20). *Meta: Systemic Censorship of Palestine Content | Human Rights Watch*. https://www.hrw.org/news/2023/12/20/meta-systemic-censorship-palestine-content

Jiménez-Durán, R. (2023). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. *George J. Stigler Center for the Study of the Economy & the State Working Paper, 324*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4590147

Jones, V. Dirty world entertainment recordings LLC (United States court of appeals for the sixth circuit June 16, 2014).

Kadri, T. E., & Klonick, K. (2019). Facebook v. Sullivan: Public figures and newsworthiness in online speech. *Southern California Law Review*, *93*, 37.

Karanicolas, M. (2021). Authoritarianism as a service: India's moves to weaponize private sector content moderation with the 2021 information technology rules. *Indian Journal of Law and Technology*, *17*, 25.

Kaye, D. B. V., Zeng, J., & Wikstrom, P. (2022). *TikTok: Creativity and culture in short video*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=KWl2EAAAQBAJ&oi=fnd&pg=PT24&dq=On+other+hand,+Launched+in+2017,+TikTok+has+rapidly+emerged+as+one+of+the+fastest-growing+social+media,+facilitating+the+sharing+of+short-form+video+content+with+its+audience&ots=eoOoUgUnE4&sig=lckIG28jC_7Un7M6_dnoLXJc-mE

Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. (2021). Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, *71*, 431–478. doi: 10.1613/jair.1.12229

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

Klonick, K. (2017). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, *131*, 1598. doi: 10.2139/ssrn.2947631

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, *60*(4), 3–26. doi: 10.1177/0306396818823172

Langvardt, K. (2017). Regulating online content moderation. *Georgetown Law Journal*, *106*, 1353. doi: 10.2139/ssrn.2937787

Lanza, E., & Jackson, M. (2021). Content moderation and self-regulation mechanisms. *The Facebook oversight board and its implications for Latin America*. *Inter-American Dialogue*, *Washington, DC*. https://www.thedialogue.org/wp-content/uploads/2021/09/Facebook-Oversight-Board-Report-EN.pdf

Lefouili, Y., & Madio, L. (2022). The economics of platform liability. *European Journal of Law and Economics*, *53*(3), 319–351. doi: 10.1007/s10657-022-09728-7

Leung, J. (2022). Shortcuts and shortfalls in Meta's content moderation practices: A glimpse from its oversight board's first year of operation. *Comparative Law and Language*, *1*(2). https://teseo.unitn.it/cll/article/view/2365

Lyons, K. (2020). *Facebook reportedly bracing for US election chaos with tools designed for 'at-risk' countries—The Verge*. https://www.theverge.com/2020/10/25/21533352/facebook-us-election-chaos-at-risk-countries-trump

Marchant, E., & Stremlau, N. (2020). The changing landscape of internet shutdowns in Africa—Introduction. *International Journal of Communication*, *14*. https://ora.ox.ac.uk/objects/uuid:6a760859-7b23-44e1-b6c2-e73a335024e1

Marsoof, A., Luco, A., Tan, H., & Joty, S. (2023). Content-filtering AI systems–limitations, challenges and regulatory approaches. *Information & Communications Technology Law*, *32*(1), 64–101. doi: 10.1080/13600834.2022.2078395

Medzini, R. (2022). Enhanced self-regulation: The case of Facebook's content governance. *New Media & Society*, *24*(10), 2227–2251. doi: 10.1177/1461444821989352

Meta. (2022). *The people behind our review teams | Transparency Centre*. https://transparency.meta.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/

Meta. (2025a). *Community Standards | Transparency Centre*. https://transparency.meta.com/en-gb/policies/community-standards

Meta. (2025b, January 7). More speech and fewer mistakes. *Meta*. https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/

Mudde, C. (2019). *The far right today*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=aD25DwAAQBAJ&oi=fnd&pg=PT3&dq=Recent+events,+including+the+Christchurch+Mosque+shootings,+violence+against+the+Rohingya+population,+and+numerous+disinformation+campaigns+related+to+the+Brexit+referendum+and+the+2016+U.S.+elections,+as+well+as+Russia%E2%80%99s+invasion+of+Ukraine,+exemplify+the+pressing+need+for+regulatory+measures+addressing+speech+on+social+media.+&ots=vYiAUhvLtO&sig=TVFjfVm-aFIj2DapEDLZcchhyIE

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

Nemer, D.(2024). Mundane technologies and community informatics. *The Journal of Community Informatics*, *20*(2), Article 2. 10.15353/joci.v20i2.6111.

Nicholas, G., & Bhatia, A. (2023). *Lost in translation: Large language models in non-English content analysis*. The Center for Democracy & Technology (CDT). https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf

Nojeim, G., & Maheshwari, N. (2021). Encryption in India: Preserving the online engine of privacy, free expression, security, and economic growth. *Indian Journal of Law and Technology*, *17*, 1. doi: 10.55496/LPNZ6069

Nourooz Pour, H. (2024). Voices and values: The challenging odyssey of meta to harmonize human rights with content moderation. *International Journal of Law and Information Technology*, *32*(1), eaae009. doi: 10.1093/ijlit/eaae009

Nunziato, D. C. (2023). The digital services act and the Brussels effect on platform content moderation. *Chicago Journal of International Law*, *24*, 115. doi: 10.2139/ssrn.3887399

O'Kane, R. (2021). Meta's private speech governance and the role of the oversight board: Lessons from the board's first decisions. *Stanford Technology Law Review*, *25*, 167. doi: 10.2139/ssrn.3683663

Okkonen, L. (2024, May 21). *Meta's shareholders: We need transparency on content moderation*. Access Now. https://www.accessnow.org/meta-agm-metas-shareholders/

Ortiz, J., Young, A., Myers, M., Bedeley, R. T., Carbaugh, D., Chughtai, H., … **Gordon, S.** (2019). *Giving voice to the voiceless: The use of digital technologies by marginalized groups*. https://researchspace.auckland.ac.nz/bitstream/handle/2292/50054/Ortiz%20et%20al.%20-%202019%20-%20Giving%20Voice%20to%20the%20Voiceless%20The%20Use%20of%20Digital%20Technologies%20by%20Marginalized%20Groups.pdf?sequence=2

Oversight Board. (2022). *Protest in India against France | Oversight Board*. https://www.oversightboard.com/decision/fb-r9k87402/

Oversight Board. (2023). *Oversight Board—2023 Annual Report* [Annual Report]. https://www.oversightboard.com/wp-content/uploads/2024/06/Oversight-Board-2023-Annual-Report.pdf

Paul, K. (2024, January 31). Zuckerberg tells parents of social media victims at Senate hearing: 'I'm sorry for everything you've been through.' *The Guardian*. https://www.theguardian.com/us-news/2024/jan/31/tiktok-meta-x-congress-hearing-child-sexual-exploitation

https://doi.org/10.1017/cfl.2025.5 Published online by Cambridge University Press

Poushter, J. (2024, March 22). WhatsApp and Facebook dominate the social media landscape in middle-income nations. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/03/22/whatsapp-and-facebook-dominate-the-social-media-landscape-in-middle-income-nations/

Quintais, J. P., Appelman, N., & Fathaigh, R. Ó. (2023). Using terms and conditions to apply fundamental rights to content moderation. *German Law Journal*, *24*(5), 881–911. doi: 10.1017/glj.2023.46

Ranchordás, S. (2019). Public values, private regulators: Between regulation and reputation in the sharing economy. *The Law & Ethics of Human Rights*, *13*(2), 203–237. doi: 10.1515/lehr-2019-2005

Reuber, A. R., & Fischer, E. (2022). Relying on the engagement of others: A review of the governance choices facing social media platform start-ups. *International Small Business Journal: Researching Entrepreneurship*, *40*(1), 3–22. doi: 10.1177/02662426211050509

Rowe, J. (2022, March 2). Marginalised languages and the content moderation challenge. *Global Partners Digital*. https://www.gp-digital.org/marginalised-languages-and-the-content-moderation-challenge/

Ruijgrok, K. (2022). The authoritarian practice of issuing internet shutdowns in India: The Bharatiya Janata Party's direct and indirect responsibility. *Democratization*, *29*(4), 611–633. doi: 10.1080/13510347.2021.1993826

Sander, B. (2019). Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation. *Fordham International Law Journal*, *43*, 939. doi: 10.2139/ssrn.3301163

San Martín, P. (2023). Meta's oversight board: Challenges of content moderation on the Internet. *Erasmus Law Review*, *16*, 124. doi: 10.2139/ssrn.3864172.

Schmitz, R. M., Coley, J. S., Thomas, C., & Ramirez, A. (2022). The cyber power of marginalized identities: Intersectional strategies of online LGBTQ+ Latinx activism. *Feminist Media Studies*, *22*(2), 271–290. doi: 10.1080/14680777.2020.1786430

Seering, J., Wang, T., Yoon, J., & Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society*, *21*(7), 1417–1443. doi: 10.1177/1461444818821316

Shah, N. (2021). Digital infrastructure, liminality, and world-making via Asia|(dis) information blackouts: politics and practices of internet shutdowns. *International Journal of Communication*, *15*, 17. doi: 10.2139/ssrn.3751273

Shahid, F., & Vashistha, A. (2023). Decolonizing content moderation: does uniform global community standard resemble utopian equality or western power hegemony? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. doi: 10.1145/3544548.3581538

Shreya Singhal v. *Union of India*, No. AIR 2015 SC 1523 (Supreme Court of India March 24, 2015).

Stieglitz, S., & Dang-Xuan, L. (2013). Social media and political communication: A social media analytics framework. *Social Network Analysis and Mining*, *3*, 1277–1291. doi: 10.1007/s13278-013-0074-2

Takhshid, Z. (2021). Regulating social media in the Global South. *Vanderbilt Journal of Entertainment and Technology Law*, *24*, 1. doi: 10.2139/ssrn.3593516

Thacker, J. (2023). Does content moderation cultivate virtue online? Toward a vision of the digital public square rooted in free speech and religious freedom. *Journal of Christian Legal Thought*, *13*, 43. doi: 10.2139/ssrn.3897440

Vaccaro, K., Xiao, Z., Hamilton, K., & Karahalios, K. (2021). Contestability for content moderation. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–28. doi: 10.1145/3476059

Waldron, P., & Ann, S. C. (2023). *One-size-fits-all content moderation fails the Global South* |. Cornell Chronicle. https://news.cornell.edu/stories/2023/04/one-size-fits-all-content-moderation-fails-global-south

Zeran v. *America Online, Inc*, No. 97-1523 (United States Court of Appeals for the Fourth Circuit November 12, 1997).

Soorya Balendra is an early-career scholar and an O'Brien Graduate Fellow at the Center for Human Rights and Legal Pluralism (CHRLP) at McGill University, where he researches the impact of social media regulatory measures on online speech and expression. Before joining McGill, he was a Lecturer in Law at General Sir John Kotelawala Defence University and held academic positions at the University of Jaffna, Sri Lanka. He was a nonresident fellow at the Information Society Law Center (ISLC) at the University of Milan (2021–2022) and worked as a Research Associate for the Digital Democracy Project at Democracy Reporting International (DRI) in 2022. His Master's research at McGill is funded by the prestigious O'Brien Graduate Fellowship. Soorya has authored national and international research publications on social media regulations, digital authoritarianism, and free speech online. He holds a Bachelor of Laws from the University of Jaffna, Sri Lanka, and is currently completing his LL.M (Thesis) at McGill. His latest book, Free Speech in the Puzzle of Content Regulation, was published by Springer in December 2024.

Cite this article: Balendra S. (2025). Meta's AI moderation and free speech: Ongoing challenges in the Global South. *Cambridge Forum on AI: Law and Governance* **1**, e21, 1–19. https://doi.org/10.1017/cfl.2025.5