Received: 13 January 2023 | Revised: 8 June 2023 | Accepted: 19 June 2023

DOI: 10.1111/1758-5899.13243

POLICY INSIGHTS

Inequalities and content moderation

Giovanni De Gregorio[1] | **Nicole Stremlau**[2,3]

[1]Católica Global School of Law, Lisbon, Portugal

[2]University of Oxford, Oxford, UK

[3]University of Johannesburg, Johannesburg, South Africa

Correspondence
Nicole Stremlau, University of Oxford, Oxford, UK.
Email: nicole.stremlau@csls.ox.ac.uk

Funding information
European Research Council, Grant/Award Number: 716686

Abstract

As the harms of hate speech, mis/disinformation and incitement to violence on social media have become increasingly apparent, calls for regulation have accelerated. Most of these debates have centred around the needs and concerns of large markets such as the EU and the United States, or the aggressive approach countries such as Russia and China adopt to regulate online content. Our focus in this article is with the rest, the smaller markets at the periphery of the advertising industry, and the deep inequalities that current approaches to content moderation perpetuate. We outline the depth of the unequal practice of moderation, particularly across Africa, and explore the underlying political and economic factors driving this gap. While recognizing content moderation has many limitations, we conclude by underlining potential approaches to increase oversight in content moderation.

1 | INTRODUCTION

Large social media companies frequently proclaim their ability to diminish the spatial divide between individuals and nations, effacing boundaries with a utopian vision characteristic of Silicon Valley, encapsulated in Facebook's mission to "bring the world closer together" (Facebook, 2021; Zuckerberg, 2021). This grandiose, and increasingly broken save-the-world mission, not only obfuscates the potential harms these large tech companies may cause, but it often conceals entrenched inequalities, particularly in relation to aligning content moderation with international human rights standards. When the relative swift moderation of content related to the Christchurch shooting in New Zealand is juxtaposed with the negligence towards the role of social media in genocide in Myanmar in 2018 or, more recently, in the conflict in Ethiopia, the conspicuous geographical bias in content moderation is evident.

This disparity is also reflected in the divergent debates about regulatory strategies to address the challenges of content moderation. In contrast with the intensive regulatory approach of the European Union to address concerns of social media through far-reaching instruments such as the Digital Services Act, debate in the United States has largely centred on the role of the First Amendment and self-regulation by the tech companies themselves. The chasm between debates about social media regulation across the Atlantic draws attention to just how challenging this issue is to address, and even more so for countries with very different political structures, many of which are in the global south. These regions often find themselves marginalized in current regulatory discussions, even as the global proliferation of harmful speech online is raising questions about the responsibility, and the ability, of social media companies to effectively tackle these challenges.

Many countries in Africa and Asia have been regarded by US social media companies, international advocacy groups, or governments in the global north, as unable to be trusted with interventionist policies (including for economic policy or laws) or in this case the regulation of online platforms, which may be seen as a proxy for censoring speech. In some cases, these concerns may be justified. Efforts by African governments to regulate social media, address online hate or counter misinformation during election periods have been interpreted as veering towards censorship, stifling dissent or suppressing protests. Many of the current legislative initiatives proposed by African governments to address harmful online content have been criticized as threats to freedom of expression, thereby placing additional responsibilities

on social media companies to more effectively manage harmful content on their platforms in such contexts.

This article proceeds by outlining the challenges facing content moderation in regions that are frequently disregarded in global debates on this transnational issue, with a focus on Africa. By examining markets that are seen as marginal from a profitability perspective, we underscore the inequalities driving contemporary approaches to moderation, including the socio-technical dimensions of content moderation, as well as the political and economic forces shaping these approaches. In certain parts of the world, this situation is exacerbated by social media companies' neglect and the insufficient allocation of resources – both in terms of human moderators and artificial intelligence systems – and significant barriers faced by governments in negotiating regulatory frameworks for content moderation. Therefore, our emphasis lies on the exacerbation of inequalities and the difficulties in implementing regulatory reforms to address the challenges of content moderation in these regions. We conclude by highlighting the tension between national efforts to address the inherent weaknesses, or limits, of content moderation and the challenges of ensuring that corporations take more substantial actions, particularly concerning mechanisms to increase oversight on content moderation.

2 | INEQUALITIES BETWEEN PRIORITY AND MARGINAL MARKETS

Inequalities in content moderation are being driven by a lack of financial incentives for social media companies to invest time, skills and resources into complex, yet relatively unprofitable markets. While the ire and frustration coming from countries such as New Zealand, Germany or France towards Facebook or YouTube's inability to control disinformation and hate speech have found more engagement at the headquarters of social media companies, the same attention has not extended to poorer countries and those that typically resort to Internet shutdowns have far less leverage over the large US companies. The GDP of a country like Burundi is approximately 3 billion USD while Meta has been valued at more than 300 times that with estimations up to 1 trillion USD, before the metaverse launch (Picchi, 2022; Worldometer, 2017), and the advertising markets in Africa, which feed the business model of social media companies, are incomparable with that of United States or the EU.

These inequalities are also underlined by the small physical footprint of social media companies in different areas of the world, and particularly in Africa with few offices and staff, if they have any at all. Facebook, for example, opened its first office on the continent in 2015 in South Africa followed by an office in Nigeria in 2021. Furthermore, social media companies do not translate their community guidelines in all the languages for the communities they support and market their products to. For Facebook, community standards were translated in only 41 out of the 111 languages supported and the rules are translated depending on criteria defined by Facebook such as whether a language has a critical mass of usage (Facebook, 2019; Fick & Dave, 2019). Notwithstanding the rhetoric by social media companies, which often argue they are representing a global community and enhancing free speech transnationally (Zuckerberg, 2017), their primary objective remains profit-oriented and not to act in the public interest. Yet content moderation is an indispensable and central role in what they do (Gillespie, 2018), considering their business model based on advertising revenues. Therefore, given these severe inequalities, it is not surprising that complaints from countries in Africa have little traction within the corporate board of directors.

Recently, whistleblower accounts have provided insights into the internal processes driving content moderation practices at social media companies, such as Facebook and Twitter, shedding light on the potential harms. Revelations by Sophie Zhang and Frances Haugen brought attention to the biased response of companies to political manipulation allegations based on the strategic political or economic importance of a country to US foreign policy and raised legitimate concerns about the influence of national foreign policy on content moderation, undermining principles of freedom of expression. According to Zhang's account, Facebook typically responded quickly to allegations or reports of political manipulation in countries such as Taiwan, South Korea and Poland, which are of strategic political importance to the United States, or to countries such as Germany or France that offer large markets and potentially significant economic consequences for non-compliance with domestic rules on hate speech or disinformation (Wong, 2021).

This is in contrast with countries or issues that have less apparent political consequences for the United States, or are smaller markets with governments that have little sway over a large foreign company. There was a case, for example, in Honduras involving the president, Juan Hernandez, and the elections. Facebook did not react for more than a year to a large, coordinated disinformation programme. This reflected its priority system for protecting political discourse and elections. In a leaked email, a Facebook executive argued "We have literally hundreds or thousands of types of abuse That's why we should start from the end (top countries, top priority areas, things driving prevalence, etc.) and try to somewhat work our way down" (Wong & Ernst, 2021). Another email indicated that the priorities for investigating this type of interference should focus on "the US/western Europe and foreign adversaries such as Russia/Iran/etc." (ibid). This echoes others

that have argued how social media are influenced by a certain cultural ideology, or their Silicon values (York, 2021), which often fails to understand the nuances of other areas of the world. The influence of political priorities driving content moderation are not isolated to US companies. A leaked excerpt of TikTok's moderation rules by a whistleblower also highlighted questions about the sway of China's foreign policy priorities over social media policies affecting users outside of China (Chen, 2020), which involved politically motivated moderation that could be seen as challenging principles of free expression and the sharing of user data with the Chinese government (Milmo, 2022).

The role of corporate identity and its susceptibility to reputational damage associated with offline harms seems to exert limited influence in compelling firms to address content moderation challenges in non-priority markets. For instance, online platforms aim to maintain control over the enforcement of their community guidelines and agreements to demonstrate that they act responsibly by complying with government requests relating to specific content categories, such as expressions associated with terrorism. Only some of the most egregious cases, which have had intensive international engagement, such as Myanmar where a UN Commission publicly argued social media was "fanning the flames of genocide" led Facebook to revise its community guidelines and employ Burmese-speaking staff (McPherson, 2018). Despite such incidents, political pressure has not substantially altered social media companies' approach to content moderation in less lucrative markets.

Language diversity is a critical factor in perpetuating inequalities. The vast range of dialects and vernacular languages inherently challenges the development of universal systems for detecting hate speech, thereby diminishing the efficacy of content moderation. This presents a scalability issue for social media platforms. AI demands extensive datasets of, for instance, text, images and user reactions, to detect patterns and categorize content for subsequent removal, blocking or shadow-banning (Myers West, 2018). As social media companies have increasingly turned to automated decision-making due to the vast amount of content (Gillespie, 2020), they still have little incentive to invest resources to address the challenges of language diversity in Africa. The performance of content moderation is firmly connected to the capacity of AI to detect and understand the context of expressions circulating on social media (Caplan, 2018). While language processing tools are more developed to detect hate in certain languages, they fail to accurately interpret the majority of idioms.

The lack of language training is also confirmed by Facebook, which reported in 2019 that its system could moderate hate speech in about 40 languages (Perrigo, 2019). By 2021, this increased to around 70 languages, and the company claims that its machine learning-based language translation engine now covers 200 languages (Fried, 2022). The extent to which it has depth of capacity in these languages, or their direct application to content moderation, is unclear, particularly in linguistically diverse countries like Ethiopia which itself has more than 70 languages, from Somali to Afar to Wolaytta. As reported by Fick and Dave (2019), the Facebook ban on glorifying organizations or terrorism has not been enough to remove posts in Somali celebrating the triumphs of the al-Shabaab militant group, with such content remaining online for months. For many African languages, like Somali, human moderators have been at the frontline of policing hate speech (Barrett, 2020). While algorithmic technologies were considered a solution to overcome human biases, there is a reverse trend in relying on human moderators to fix artificial intelligence biases (O'Neil, 2021). Indeed, even if human judgement tends to be biased, there is a paradox to use algorithmic technologies to cope with human bias (Ajunwa, 2020).

The Myanmar genocide underlined this challenge (Stecklow, 2018). The spread of hate speech on Facebook has been attributed to supporting ethnic cleansing in Myanmar, a situation which largely went unchecked due to inadequate moderation tools and personnel capable of understand the local language. While Facebook has since employed Burmese speakers as human moderators to compile a data set of hateful and violent expressions, the international pressure has also led to overreactions with speech that might be considered legitimate, banned (Sablosky, 2021).

Considerable technical constraints exist in training algorithmic technologies to detect harmful content in a wide range of languages. Martinus and Abbott (2019) underlined some of the primary obstacles in machine translation for African languages such as limited availability and discoverability. The lack of language training also comes from the difficulty in accessing resources in uncommon languages. For instance, Amharic and Somali, along with other African languages, are low resource languages which poses a substantial challenge to machine learning systems. Even though certain data repositories in African languages do exist (Marivate, 2020), restrictions in data and publication accessibility are one of the primary barriers for the language training of artificial intelligence technologies (Braun & Ong, 2014). As underlined by Guy Rosen, Facebook vice president, machine learning needs to process vast amounts of data for training and the scarcity of text in other languages compounds the issue (Fick & Dave, 2019). Africa is a mosaic of languages, with the number of living languages and dialects estimated to be in the thousands.

The rapid increase in African internet users, combined with the continent's enormous linguistic diversity, has spurred efforts to collect data and develop

methodologies to extend research in these languages (these efforts, however, remain modest compared with languages in the Global North; Kann et al., 2019; Petrollino et al., 2019). Taking Ethiopia as an example of a nation afflicted by violent conflicts and hate speech, not only in Tigray but in other regions as well, there are few computational linguistic resources that have been developed for Ethiopia's wide range of languages. Mossie and Wang (2018) underlined that still there is more work to do in terms of expanding the dataset and the statistical significance, improve the contextualisation of hate based on different criteria (e.g., ethnicity), access the information provided by social media and add other sources to improve the space for this under-resourced language. Indeed, even if Amharic is prevalent, is not the only language spoken in Ethiopia. While Amharic is a language supported by Facebook, moderation has failed to effectively to deal with the situation. Posts in Amharic have been used for ethnic clashes attacking the Oromo and Tigray (Wilmot, 2020). A similar situation exists in other countries. Shikali and Mokhosi (2020) have focused on providing a data set for Swahili, which is spoken in Eastern Africa, to enhance the possibility to detect content in this language which is spoken by more than 15 million people as native speakers and 82 as a secondary language. And in Western Africa, Doumbouya et al. (2021) worked on the first speech recognition models for Maninka, Pular and Susu. In this case, the data have been based on radio broadcasting archives which include abundant data in low-resource languages. In South Africa, De Vries et al. (2014) collected hundreds of hours of speech in all 11 languages. Other works have focused on multi-lingual speakers (Modipa et al., 2013) and using soap-opera data speech (van der Westhuizen & Niesler, 2018). There has been active work on Setswana (Marivate et al., 2020) and isiXhosa (Packham & Suleman, 2015).

Addressing content without a shared concept of harmfulness, like hate speech, presents formidable challenges in the development of detection tools. While images or videos often transcend cultural and linguistic barriers, even if semantically different, the ubiquitous presence of text on post or comments makes the detection of harmful content harder. This problem extends to images or videos with embedded text which include hate speech or violent messages. Even assuming that AI can detect some forms of hate speech in video or images, there remain uncertainties about whether these systems can distinguish context. Given their transnational deployment, detection technologies tend to function as broad tools rather than precise instruments, leading to potential false positives such as over-censorship, or false negatives when content is kept online despite a lack of compliance with community guidelines. Machine learning systems cannot always understand the semantics of certain content and accommodate language changes over time. Understanding context means dealing with cultural and historical information in different geographical locations.

The lack of language processing tools in local languages precludes the reliance on opinion mining techniques to analyse sentiments, emotions and attitudes (Liu, 2020). Given the absence of precise tools understanding local languages, employing artificial intelligence technologies to process content such as emotions or reactions- commonly referred to as emotional AI (LaGrandeur, 2015; McStay, 2018; Stark & Hoey, 2021) – raises concerns about how content moderation can promote hate and disappointment, thus leading to radicalisation or polarization (Allan, 2013). This issue is particularly pertinent when considering the potential role of echo chambers (Pariser, 2011; Sunstein, 2018), especially in areas of the world where conflicts are being driven by social and religious tension.

Moreover, the lack of language and context training also suggests that artificial intelligence technologies cannot accommodate the multiplicity of definitions and sensitivity of hate speech in Africa (Gagliardone et al., 2016; Pohjonen, 2019). Even within a legal framework, there no unified legal notion of hate speech (Brown & Sinclair, 2019; Heinze, 2016). The fragmentation of legal definitions does not encapsulate the full scope of hate speech beyond the legal sphere, highlighting the importance of distinguishing between hate speech as a legal concept and its use in every day discourse (Brown, 2017a, 2017b). Consequently, the meaning of hate speech is shaped by social norms and cultural values, not just by those with legal expertise.

Computer scientists working on programming and developing algorithmic technologies to moderate content have put forth varying definitions of hate speech. Natural language processing has concentrated on hate speech and toxic content (Burnap & Williams, 2015; Davidson et al., 2017) while developing tools for detection (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Schmidt & Wiegand, 2017; Waqas et al., 2019). Most of these technologies are programmed, developed and trained in the context of the English or Western languages (Del Vigna et al., 2017; Tulkens et al., 2016), primarily due to the greater availability of data in these languages to train algorithmic technologies. However, while these languages are diffused on a global scale, hate or disinformation is not only a matter for some languages or areas.

While social media companies are doing more in terms of hiring human moderators and creating data sets of hate speech, as seen in the aftermath of the Rohingya genocide (Perrigo, 2019), there is still little access to more granular information about the position of local moderators, the building of data sets or the improvements in algorithmic decision-making. Recently

in the case of Azerbaijan where Facebook posts have been implicated in the war with Armenia over the disputed territory, a whistleblower highlighted that 'Azerbaijan fell into a gap: neither the eastern European nor the Middle Eastern policy teams claimed responsibility for it, and no operations staff – either full-time or contract – spoke Azeri' (Wong & Harding, 2021).

Even when expertise in language and context is accessible through human moderators, the issue of biases is far from completely resolved. Programmers and moderators play a significant role in translating human biases into machine biases within artificial intelligence technologies (Barocas & Selbst, 2016; Pasquale, 2015). It is important to emphasize that algorithmic tools are not inherently neutral. Consequently, when human moderators are supplanted by algorithmic technologies, the risks associated with bias and errors remain prominent. These technologies are developed by humans and trained using datasets that may include biases. Such datasets can reflect societal biases, thereby mirroring the prevailing inequalities within our society (Condliffe, 2019). For instance, Twitter's AI-generated images which displayed a preference for Caucasian faces over those of people of colour (Lyons, 2020).

Addressing this issue, some researchers have suggested relying on balanced classifiers and metrics specifically tailored for subgroups, such as dark females (Buolamwini & Gebru, 2018). Nevertheless, existing social biases are further entrenched by algorithmic detection tools primarily developed in certain languages or by human annotators (Binns et al., 2017). While it is well-known how search engines provide a channel for discriminating against marginalized groups (Eubanks, 2018; Hankerson et al., 2016; O'Neil, 2016), research has also shown that communities speaking dialects or less common languages are subject to over-censorship because detecting tools are predominantly developed in English (Sap et al., 2019). This situation underscores how the process of content moderation is not only affected by the lack of incentives for social media platforms but also by the deployment of automated decision-making technologies.

3 | LOCAL AND INTERNATIONAL REPONSES TO INEQUALITIES

The lack of incentives for social media platforms to improve their content moderation process in Africa, coupled with insufficient language training, are not the only sources of inequality. Against a backdrop of opaque governance structures and weak responses by social media companies, a variety of actors, from non-governmental organizations to various public authorities worldwide have attempted to mitigate the harm wrought by the proliferation of hate and disinformation online. The spread of online hate and disinformation has also been used to buttress arguments for new information control measures. This section will focus on how national governments are responding to this real, or perceived, injustice.

In the context of the United States, social media platforms have been seem by some to be "governors" (Klonick 2018) that retain the freedom to decide on content moderation strategies and to establish appropriate mechanisms to deal with the spread of objectionable content. Meanwhile in Europe, a new regulatory phase of content moderation is emerging with the Digital Services Act. It aims to modernize the governing online intermediaries while remaining rooted to exemption of liability of online platforms for hosting unlawful content envisaged by the e-Commerce Directive. The Digital Services Act introduces a new set of procedures aiming to increase the level of accountability in content moderation, including due diligence and transparency requirements while providing redress mechanisms for users. In other words, without regulating content, it requires that online platforms comply with procedural safeguards in content moderation.

However, governments outside of the EU have exhibited varied responses, particularly in less affluent and less geopolitically influential countries. While the digital spaces provided by social media platforms have bolstered access to diverse information online, promoting a plurality of voices and the exchange of opinions, the potential use of these technologies to contest central authorities and spread disinformation has encouraged governments to censor online speech or use social media as instrument of surveillance (Morozov, 2012; Zittrain et al., 2017).

Within the African context, governmental accusations that platforms are disseminating hate and disinformation have not led to a regulatory response favouring transparency and accountability. Instead, these accusations have generally led to amplified efforts to criminalize certain forms of speech on social media (Kshetri, 2019; Olewe, 2018), while blaming companies for spreading violence (Wilmot, 2020). This situation has furnished governments with a pretext to introduce controversial laws criminalizing some social media activity. In the case of Nigeria, for example, two bills have been proposed to increase government powers to shut down the internet, punish government critics and sanction hate speech with capital punishment (Abdulrauf, 2019). Among other strategies, Uganda's political approach to online speech restriction has been unusual introducing taxes for the use of Internet services and data including Facebook, Twitter, Skype and WhatsApp (Mwesigwa, 2021).

Alongside legislation, African governments have also resorted to blocking content (Giles & Mwai, 2021). Yet, according to Google's transparency report on government order to remove content in the first half of 2022, only a few African countries such as Kenya,

South Africa and Angola requested Google to remove content. These requests, fewer than 30 in number for Africa (Google, 2022), highlight a significant gap when compared with the data of countries like the United States, Germany or Russia, whose requests are in the tens of thousands.

Some countries also do not necessarily subscribe to this game. Instead, they chose to rely on tactics other than reporting hate speech. Computational propaganda is an increasingly common tool employed by governments, including Ethiopia, Rwanda and Sudan (Bradshaw & Howard, 2019). An expose by *The Guardian* revealed internal documentation showing how Facebook handled more than 30 cases across 25 countries of politically manipulative behaviour that was proactively detected by company staff. This was likely just a small sample, yet the company declined to take action on these cases, citing both priorities and capacity constraints (Wong, 2021).

This trend also suggests that African governments have relied more on direct censorship rather than cooperating with social media companies to remove content. The proliferation of hate speech on social media has become a primary justification for the increasing governmental use of internet shutdowns. These measures can range from throttling internet speed to the point if making it practically unusable, to completely switching it off (De Gregorio & Stremlau, 2020). Whereas only a few years ago such forms of censorship would be seen as a grave violation of freedom of expression, increasingly they are understood as one of the few mechanisms available for addressing online speech and offline harms in a moment of crisis. Although empirical evidence demonstrating the effectiveness of these practices in combatting misinformation and hate speech is scant (Kingsley, 2019), shutdowns have been employed to curb online discourse, especially during election periods.

The escalation of internet shutdowns also reflects the frustration on the part of some governments due to their inability to intervene in the governance of online platforms that are often in another jurisdiction, on another continent. In the absence of concerted cooperation with companies, shutting down the entire network or specific digital spaces has become increasingly popular. While social media try to extend their policy of moderation on a global scale, in some cases they try to comply with local laws and accommodate orders by governments. The case of the Indian request during Covid-19 is such an example, where the companies collaborated with the government to moderate content (Clarke & Swindells, 2021). But such efforts often depend on the financial incentives, including whether they have a footprint in the country and the level of advertising revenues.

There are also cases where platforms resist government demands. A notable example is the case of the "Innocence of Muslims" video, which incited global protests and the attack on the US consulate in Libya. The US government requested YouTube to review the video in light of its community guidelines, but the platform opted to keep the content online (Gerstein, 2012). In other cases, social media have independently removed content by governments. Twitter took down a tweet from Nigerian president Muhammadu Buhari that threatened to punish regional secessionists. This decision, however, incited retaliation from the Nigerian government which subsequently blocked Twitter. Similarly, in Uganda, the government shut down the Internet after the decisions by Facebook and Twitter to suspend the accounts of the ruling party for coordinated inauthentic behaviour around the 2021 elections. The African Commission on Human and People's Rights issued a declaration on freedom of expression and access to information in Africa (2019), calling for increasing multistakeholder cooperation between public and private actors to adopt a strict human rights approach when designing content moderation. The conflicts between governments and social media have not resulted in enhanced user protections but an increase in internet shutdowns. This oscillation between resistance and collaboration is firmly intertwined with the opaque nature of content moderation.

4 | OVERSEEING CONTENT MODERATION

In regions with less economic incentives for major social media companies, artificial intelligence is unlikely to offer a radical remedy to redress the deficit of effective moderation as some have hoped, at least in the near future. This raises difficult questions about how to redress this disparity. While content moderation reflects broader power imbalances on a global scale, considering emerging strategies aiming to foster spaces for conducive more effective content moderation might be constructive. These strategies include bottom-up, or grassroots and top-down methods, along with national and international approaches.

Beginning with bottom-up alternatives, users currently have a tangential role in either setting the rules or aiding processes of content moderation. They are primarily viewed by companies as sources from which to extract data that can help to offer targeted services to attract advertising revenue. Users trade some of their rights to comply with conditions determined through a top-down approach driven by commercial interests. At the same time, users are vital components of the content moderation puzzle, given social media companies depend- to varying degrees- on users to flag or report content (Crawford & Gillespie, 2016). Artificial intelligence technologies are also trained based on users' reporting of harmful content. Thus, understanding how

users identify, react, engage and report online hate speech is critical to discerning both current and potential opportunities for AI systems to detect certain forms of harm. Inaccessibility to reporting online hate speech marginalizes users in some regions, thereby reducing the benefits of content moderation and the safeguards established by social media, such as redress mechanisms.

Approaches such as the Meta Oversight Board could potentially contribute to revealing the pitfalls of content moderation and unpacking the inequality in areas that are peripheral markets for large corporations, as recently attempted by an Ethiopian case that was referred to the Board (Meta Oversight Board, 2022). However, the scale of these initiatives is modest, often devoid of context or relevance for countries in the Global South, and may form only a part of a broader approach of oversight that should involve more organizations monitoring online harm. In the case of the Oversight Board, user participation is confined to reporting content to the Board and sending comments on specific cases. Moreover, according to the Board's reports, cases from Africa account for just 1% of those received.

Civil society organizations may also play a role. The establishment of 'harm-checking' organizations would enhance oversight, particularly on online harm. Social media spaces have been exploited for spreading online hate and disinformation and have amplified the reach of content and engagement on a global scale. Analogous to fact-checkers, these organizations could augment monitoring over online hate speech and promote local initiatives for digital skill development. However, in this case, these organizations are likely to face questions about their independence, particularly concerning their funding and challenges related to language diversity, limited capacity and access to enough data by social media companies to capture the wide range of content being produced.

From the perspective of top-down approaches, many governments in the global south possess limited capabilities to tackle content moderation challenges. The increasing reliance on Internet shutdowns exemplifies how some governments have addressed online harm, often using content moderation challenges to justify these measures to limit access. In some cases, social media has been exploited for political purposes, particularly to spread disinformation and propaganda, while, at the same time, criminalizing this content. While internet shutdowns do pose a challenge to social media companies because their products are restricted and, on a more ideological level, shutdowns are in opposition to their beliefs of global connectivity, it is unlikely that a top-down approach to content moderation would result from incentives provided by social media companies.

External trends in the regulation of content moderation may exert influence on these smaller markets. The potential global impact not only of controversies but also the EU model can play a critical role, particularly looking at the new procedural safeguards in content moderation introduced by the Digital Services Act. This model is not based on content regulation but on the proceduralisation of content moderation. Rather than criminalizing hate speech, the Digital Services Act introduces procedural safeguards that make social media more transparent and accountable. Content moderation is fundamentally driven by business logic and is a matter of scale. If social media platforms are required to comply with EU standards, it is probable that they will make their processes more transparent even beyond the EU. Similarly, governments could also be incentivized to adopt the same approach, necessitating greater transparency and accountability from social media platforms. Specifically, the potential to obtain more information about content removal and flagging in peripheral areas, as well as the right to access data for research, would be important steps further to study online harm.

However, these external influences can provide sources of legitimation for governments to regulate social media, thereby creating spaces to manipulate procedural safeguards for political purposes. The adoption of Germany's NetzDG has become an example for countries around the world which have transformed this legislation into an instrument of information control (Mchangama & Fiss, 2019). This is part of a broader history of legal transplants that is particularly concerning when they move across different legal systems, particularly from the west to other contexts.

The oversight of content moderation can also be reinforced through intergovernmental agencies such as the United Nations. Longstanding debates persist regarding the circumstances and methods of intervening in cases of hate speech and mass atrocities, including the balance between freedom of expression and the responsibility to protect. The formalization and institutionalization of norms and processes for intervening in information systems, also known as 'information interventions', could be achieved through the establishment of an 'Information Intervention Council'. As a body responsible for monitoring online harm, it could be a first step to increase oversight (De Gregorio & Stremlau, 2021). This council would be tasked with conducting research, establishing guidelines for content moderation in areas that are often neglected by corporations, and supporting the formulation of guidelines to encourage the private sector's compliance with specific standards. There are, for example, 'due-diligence guidelines' promoting a specific code of conduct for companies operating in the import, processing and sale areas of the minerals extracted in places such as the Democratic Republic of Congo, aiming to reduce exacerbating the conflict in the Eastern part of the country (Security Council,

Resolution 1952, 2010). In addition, the UN Security Council has also advocated new public-private partnerships to address global challenges like terrorism, particularly highlighting the role of social media (Resolution 2354, 2017). Establishing an international framework could also help mitigate more extreme information intervention measures—such as internet shutdowns—that are frequently implemented in an ad hoc manner and seldom through formal policy or legal channels.

The inequalities, and disparities, of content moderation cannot be rectified through a single approach, but rather a hybrid strategy based on bottom-up, top-down and international methods. To enhance oversight on content moderation processes, it is crucial to create a system that reflects the diversity and influences of multiple stakeholders. This is not a call for revitalizing multi-stakeholderism, but rather an emphasis on the contribution of diverse elements to mitigate the imbalances created by the dominance of large social media companies.

ACKNOWLEDGEMENTS

This research has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 716686, ConflictNet).

REFERENCES

Abdulrauf, A. (2019) Nigeria bill aims at punishing hate speech with death. *Deutsche Welle* (26 November). Available from: https://www.dw.com/en/nigeria-bill-aims-at-punishing-hate-speech-with-death/a-51419750 [Accessed 13th July 2023].

Ajunwa, I. (2020) *The paradox of automation as anti-bias intervention, 41 Cardozo, L*. Available at SSRN 2746078.

Allan, J. (2013) Hate speech law and disagreement. *Constitutional Commentary*, 29, 59–79.

Barocas, S. & Selbst, A.D. (2016) Big data's disparate impact. *California Law Review*, 104, 671.

Barrett, P. (2020) Who moderates the social media giants? A call to end outsourcing. *NYU Center for Business and Human Rights*, 1–32. Available from: https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf [Accessed 13th July 2023].

Binns, R., Veale, M., Kleek, M.V. & Shadbolt, N. (2017) Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: *International conference on social informatics*. Cham: Springer, pp. 405–415. Available from: https://doi.org/10.1007/978-3-319-67256-4_32

Bradshaw, S. & Howard, P.N. (2019) *The global disinformation order: 2019 global inventory of organised social media manipulation*. Available from: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf [Accessed 13th July 2023].

Braun, M.L. & Ong, C.S. (2014) Open science in machine learnin. In: Stodden, V., Leisch, F. & Peng, R.D. (Eds.) *Implementing reproducible research*. New York: Chapman and Hall/CRC, pp. 343.

Brown, A. (2017a) What is hate speech? Part 1: the myth of hate. *Law and Philosophy*, 36(4), 419–468.

Brown, A. (2017b) What is hate speech? Part 1: the myth of hate. *Law and Philosophy*, 36, 561–613.

Brown, A. & Sinclair, A. (2019) *The politics of hate speech laws*. Oxford: Routledge.

Buolamwini, J. & Gebru, T. (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency*. PMLR, pp. 77–91. Available from: http://proceedings.mlr.press/v81/buolamwini18a.html [Accessed 13th July 2023].

Burnap, P. & Williams, M.L. (2015) Cyber hate speech on twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 223–242.

Caplan, R. (2018) Content or context moderation? Artisanal, community-reliant, and industrial approaches. *Data & Society*, 14. Available from: https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf [Accessed 13th July 2023].

Chen, A. (2019) A leaked excerpt of TikTok moderation rules shows how political content gets buried. *MIT Technology Review*, 25 November.

Clarke, L. & Swindells, K. (2021) How social media companies help authoritarian governments censor the internet. *Newstatesman*, June, 9. Available from: https://www.newstatesman.com/business/companies/2021/06/how-social-media-companies-help-authoritarian-governments-censor-internet [Accessed 13th July 2023].

Condliffe, J. (2019) The week in tech: algorithmic bias is bad. Uncovering it is good. *The New York Times*. Available from: https://www.nytimes.com/2019/11/15/technology/algorithmic-ai-bias.html [Accessed 13th July 2023].

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

Davidson, T., Warmsley, D., Macy, M. & Weber, I. (2017) Automated hate speech detection and the problem of offensive language. In: *Proceedings of the international AAAI conference on web and social media*, vol. 11, No. 1, pp. 512–515.

De Gregorio, G. & Stremlau, N. (2020) Internet shutdowns and the limits of the law. *International Journal of Communication*, 14, 4224.

De Gregorio, G. & Stremlau, N. (2021) Platform governance at the periphery: moderation, shutdowns and intervention. In: Bayer, J., Holznagel, B., Korpisaari, P. & Woods, L. (Eds.) *Perspectives on platform regulation. Concepts and models of social media governance across the globe*. Nomos.

De Vries, N.J., Davel, M.H., Badenhorst, J., Basson, W.D., Wet, D. et al. (2014) A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, 56(1), 119–131.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M. & Tesconi, M. (2017) Hate me, hate me not: hate speech detection on Facebook. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pp. 86–95.

Doumbouya, M., Einstein, L. & Piech, C. (2021) Why AI needs to be able to understand all the world's languages. *Scientific American*. Available from: https://www.scientificamerican.com/article/why-ai-needs-to-be-able-to-understand-all-the-worlds-languages/ [Accessed 13th July 2023].

Eubanks, V. (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.

Facebook. (2019) *Community standards enforcement report, January–March 2019*. Available from: https://transparency.facebook.com/community-standards-enforcement#hate-speech [Accessed 13th July 2023].

Facebook. (2021) *Of the violating content we actioned for hate speech, how much did we find before people reported it?* Available from: https://transparency.fb.com/data/community-stand

ards-enforcement/hate-speech/facebook [Accessed 13th July 2023].
Fick, M. & Dave, P. (2019) Facebook's flood of languages leave it struggling to monitor content. *Reuters*, April, 23. Available from: https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW [Accessed 13th July 2023].
Fortuna, P. & Nunes, S. (2018) A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4), 1–30. Available from: https://doi.org/10.1145/3232676
Fried, I. (2022) Facebooks AI translator now works with 200 languages. *Axios*. Available from: https://www.axios.com/2022/07/06/facebook-ai-translator-200-languages [Accessed 13th July 2023].
Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M. et al. (2016) *Mechachal: online debates and elections in Ethiopia-from hate speech to engagement in social media*. Available at SSRN 2831369.
Gerstein, J. (2012) Activists troubled by white house call to YouTube. *Politico*, September, 14. Available from: https://www.politico.com/blogs/under-the-radar/2012/09/activists-troubledby-white-house-call-to-youtube-135618 [Accessed 13th July 2023].
Giles, C. & Mwai, P. (2021) Africa internet: where and how are governments blocking it. *BBC News*. Available from: https://www.bbc.com/news/world-africa-47734843 [Accessed 13th July 2023].
Gillespie, T. (2018) *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. New Heaven, CT: Yale University Press.
Gillespie, T. (2020) Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 2053951720943234. Available from: https://doi.org/10.1177/2053951720943234
Google. (2022) *Government requests to remove content*. Available from: https://transparencyreport.google.com/government-removals/government-requests/DZ [Accessed 13th July 2023].
Hankerson, D., Marshall, A.R., Booker, J., El Mimouni, H., Walker, I. & Rode, J.A. (2016) Does technology have race? In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 473–486. Available from: http://doi.org/10.1145/2851581.2892578
Heinze, E. (2016) *Hate speech and democratic citizenship*. Oxford: Oxford University Press.
Kann, K., Cho, K. & Bowman, S.R. (2019) Towards realistic practices in low-resource natural language processing: the development set. *arXiv preprint arXiv:1909.01522*. Available from: http://doi.org/10.18653/v1/d19-1329
Kingsley, P. (2019) Life in an internet shutdown: crossing borders for email and contraband SIM cards. *The New York Times*. Available from: https://www.nytimes.com/2019/09/02/world/africa/internet-shutdown-economy.html [Accessed 13th July 2023].
Klonick, K. (2018) The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6), 1599.
Kshetri, N. (2019) Cybercrime and cybersecurity in Africa. *Journal of Global Information Technology Management*, 22(2), 77–81. Available from: https://doi.org/10.1080/1097198X.2019.1603527
LaGrandeur, K. (2015) Emotion, artificial intelligence, and ethics. In: *Beyond artificial intelligence*. Cham: Springer, pp. 97–109. Available from: https://doi.org/10.1007/978-3-319-09668-1_7
Liu, B. (2020) *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press.
Lyons, K. (2020) Twitter is looking into why its photo preview appears to favor white faces over black faces. *The Verge*. Available from: https://www.theverge.com/2020/9/20/21447998/twitter-photo-preview-whiteblack-faces [Accessed 13th July 2023].
MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. & Frieder, O. (2019) Hate speech detection: challenges and solutions. *PLoS One*, 14(8), e0221152. Available from: https://doi.org/10.1371/journal.pone.0221152
Marivate, V. (2020) *Why African natural language processing now? A view from South Africa# AfricaNLP*. Available from: https://mistra.org.za/wp-content/uploads/2020/11/4IR-Working-Paper_Dr-Vukosi-Marivate_20201102.pdf [Accessed 13th July 2023].
Marivate, V., Sefara, T., Chabalala, V., Makhaya, K. & Mokgonyane, T. (2020) Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi. In: *Proceedings of the first workshop on Resources for African Indigenous Languages*, 15–20.
Martinus, L. & Abbott, J.Z. (2019) A focus on neural machine translation for African languages. *arXiv preprint arXiv:1906.05685*.
Meta Oversight Board. (2022) Available from: https://www.oversightboard.com/news/592325135885870-oversightboard-upholds-meta-s-decision-in-tigray-communication-affairs-bureau-case-2022-006-fb-mr/ [Accessed 25 July 2023].
Mchangama, J. & Fiss, J. (2019) The digital Berlin Wall: how Germany (accidentally) created a prototype for global online censorship. *Justitia*. Available from: https://justitia-int.org/wp-content/uploads/2019/11/Analyse_The-Digital-Berlin-Wall-How-Germany-Accidentally-Created-a-Prototype-for-Global-Online-Censorship.pdf [Accessed 13th July 2023].
McPherson, P. (2018) Facebook says it was 'too slow' to fight hate speech in Myanmar. *Reuters*, August 16. Available from: https://www.reuters.com/article/us-myanmar-facebook-rohingya-idINKBN1L1066 [Accessed 13th July 2023].
McStay, A. (2018) *Emotional AI: the rise of empathic media*. London: Sage.
Milmo, D. (2022) TikTok's ties to China: why concerns over your data are here to stay. *The Guardian*, November 8. Available from: https://www.theguardian.com/technology/2022/nov/07/tiktoks-china-bytedance-data-concerns [Accessed 13th July 2023].
Modipa, T.I., De Wet, F. & Davel, M.H. (2013) Implications of Sepedi/English code switching for ASR systems. In: *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa*.
Morozov, E. (2012) *The net delusion: the dark side of internet freedom*. New York: PublicAffairs.
Mossie, Z. & Wang, J.H. (2018) Social network hate speech detection for Amharic language. *Computer Science & Information Technology*, 41–55. Available from: https://doi.org/10.5121/csit.2018.80604
Mwesigwa, D. (2021) Uganda abandons social media tax but slaps new levy on internet data. *Cipesa* (1 July). Available from: https://cipesa.org/2021/07/uganda-abandons-social-media-tax-but-slaps-new-levy-on-internet-data/ [Accessed 13th July 2023].
Myers West, S. (2018) Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media & Society*, 20(11), 4366–4383. Available from: https://doi.org/10.1177/1461444818773059
Olewe, D. (2018) Kenya, Uganda and Tanzania in "anti-fake news campaign". *BBC News*. Available from: https://www.bbc.com/news/world-africa-44137769 [Accessed 13th July 2023].
O'Neil, C. (2016) *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York, NY: Crown.
O'Neil, C. (2021) Facebook and twitter can't police what gets posted. Neither AI nor humans seem capable of properly moderating content. *Bloomberg*. Available from: https://www.bloomberg.com/opinion/articles/2021-02-19/facebook-and-twitter-content-moderation-is-failing [Accessed 13th July 2023].
Packham, S. & Suleman, H. (2015) Crowdsourcing a text corpus is not a game. In: *International conference on Asian digital libraries*, pp. 225–234.
Pariser, E. (2011) *The filter bubble: what the internet is hiding from you*. London: Penguin.

Pasquale, F. (2015) *The black box society: the secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Perrigo, B. (2019) Facebook says it's removing more hate speech than ever before. But there's a catch. *Time*, November, 26. Available from: https://time.com/5739688/facebook-hate-speech-languages/ [Accessed 13th July 2023].

Petrollino, S., Nyst, V., Tunde, O., Ngué Um, E., Ekpenyong, M. et al. (2019) African languages and digital humanities: challenges and solutions. In: *Digital Humanities Conference 2019*.

Picchi, A. (2022) Meta's value has plunged by $700 billion. Wall street calls it a train wreck. *Moneywatch*, October 28. Available from: https://www.cbsnews.com/news/meta-stock-down-earnings-700-billion-in-lost-value/ [Accessed 13th July 2023].

Pohjonen, M. (2019) A comparative approach to social media extreme speech: online hate speech as media commentary. *International Journal of Communication*, 13, 3088–3103.

Sablosky, J. (2021) Dangerous organizations: Facebook's content moderation decisions and ethnic visibility in Myanmar. *Media, Culture & Society*, 43(6), 1017–1042. Available from: https://doi.org/10.1177/0163443720987751

Sap, M., Card, D., Gabriel, S., Choi, Y. & Smith, N.A. (2019) The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678.

Schmidt, A. & Wiegand, M. (2017) A survey on hate speech detection using natural language processing. In: *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10.

Shikali, C.S. & Mokhosi, R. (2020) Enhancing African low-resource languages: Swahili data for language modelling. *Data in Brief*, 31, 105951.

Stark, L. & Hoey, J. (2021) The ethics of emotion in artificial intelligence systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 782–793. Available from: http://doi.org/10.1145/3442188.3445939

Stecklow, S. (2018) Why Facebook is losing the war on hate speech in Myanmar Reuters. *Reuters*, 15 August. https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ [Accessed 13th July 2023].

Sunstein, C.R. (2018) *# Republic: divided democracy in the age of social media*. Princeton, NJ: Princeton University Press.

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B. & Daelemans, W. (2016) A dictionary-based approach to racism detection in Dutch social media. *arXiv preprint arXiv:1608.08738*.

UN Resolution 2354. (2017) Available from: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N17/149/22/PDF/N1714922.pdf?OpenElement [Accessed 13th July 2023].

UN Security Council Resolution 1952. (2010) Available from: https://www.un.org/securitycouncil/s/res/1952-%282010%29 [Accessed 13th July 2023].

van der Westhuizen, E. & Niesler, T. (2018) A first south African corpus of multilingual code-switched soap opera speech. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Available from: https://aclanthology.org/L18-1451.pdf, http://www.lrec-conf.org/proceedings/lrec2018/summaries/90.html [Accessed 25 July 2023].

Waqas, A., Salminen, J., Jung, S.-g., Almerekhi, H. & Jansen, B.J. (2019) Mapping online hate: a scientometric analysis on research trends and hotspots in research on online hate. *PLoS One*, 14(9), e0222194. Available from: https://doi.org/10.1371/journal.pone.0222194

Wilmot, C. (2020) Ethiopia's cracking down in Tigray. But activists are spreading the news. *Washington Post*, 17 November. Available from: https://www.washingtonpost.com/politics/2020/11/17/ethiopias-cracking-down-tigray-activists-are-spreading-news/ [Accessed 13th July 2023].

Wong, J. & Harding, L. (2021) Facebook isn't interested in countries like ours': Azerbaijan troll network returns months after the ban. *The Guardian*. Available from: https://www.theguardian.com/technology/2021/apr/13/facebook-azerbaijan-ilham-aliyev [Accessed 13th July 2023].

Wong, J.C. (2021) Revealed: the Facebook loophole that lets world leaders deceive and harass their citizens. *The Guardian*. Available from: https://www.theguardian.com/technology/2021/apr/12/facebook-loophole-state-backed-manipulation [Accessed 13th July 2023].

Wong, J.C. & Ernst, J. (2021) Facebook knew of Honduran president's manipulation campaign – and let it continue for 11 months. *The Guardian*. Available from: https://www.theguardian.com/technology/2021/apr/13/facebook-honduras-juan-orlando-hernandez-fake-engagement [Accessed 13th July 2023].

Worldometer. (2017) *GDP by country*. Available from: https://www.worldometers.info/gdp/gdp-by-country/ [Accessed 13th July 2023].

York, J.C. (2021) *Silicon values: the future of free speech under surveillance capitalism*. London: Verso.

Zittrain, J.L., Faris, R., Noman, H., Clark, J., Tilton, C. & Morrison-Westphal, R. (2017) The shifting landscape of global internet censorship. *Berkman Klein Center Research Publication*, (2017-4), 17–38. Available from: https://dash.harvard.edu/bitstream/handle/1/33084425/The%20Shifting%20Landscape%20of%20Global%20Internet%20Censorship-%20Internet%20Monitor%202017.pdf [Accessed 13th July 2023].

Zuckerberg, M. (2017) *Building global community*. Facebook (16 February). Available from: https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/ [Accessed 13th July 2023].

Zuckerberg, M. (2021) *Bringing the world closer together*. Facebook (15 March). Available from: https://www.facebook.com/notes/mark-zuckerberg/bringing-the-world-closer-together/10154944663901634/ [Accessed 13th July 2023].

AUTHOR BIOGRAPHIES

Giovanni De Gregorio is the PLMJ Chair in Law and Technology at Católica Global School of Law and Católica Lisbon School of Law. He can be reached at gdegregorio@ucp.pt.

Nicole Stremlau is Head of the Programme in Comparative Media Law and Policy, Centre for Socio-Legal Studies at the University of Oxford where she is PI of the European Research Council ConflictNet project. She is also a Research Professor, School of Communications, University of Johannesburg. She can be reached at nicole.stremlau@csls.ox.ac.uk.

How to cite this article: De Gregorio, G. & Stremlau, N. (2023) Inequalities and content moderation. *Global Policy*, 00, 1–10. Available from: https://doi.org/10.1111/1758-5899.13243