Inside content moderation: Humans, machines and invisible work

Authors: Maurice Stenzel, Katharina Mosene & Frederik Efferenn

Who decides what we see online and what we don't? Moderating content on social media platforms is a complex process. It is shaped not only by platform-specific rules and technical infrastructures, but also by legal frameworks at national and international levels. Closely linked to this is the question of social responsibility. Content moderation goes far beyond simply deleting problematic posts: every decision directly affects platform users, determining which voices remain visible and which are silenced. The division of labour between algorithmic systems and human moderators repeatedly reaches its limits. Platform companies outsource large parts of this moderation work to countries such as the Philippines or Kenya, where people review highly distressing content under precarious conditions. Meanwhile, the algorithms and guidelines that shape their work are largely developed in the Global North. This shifting of responsibilities reproduces or even amplifies existing inequalities, for instance, along the lines of gender, origin, or ethnicity. This article presents research approaches that critically examine these power asymmetries and incorporate intersectional as well as decolonial perspectives. The goal is to make digital spaces and the way they are governed fairer and more inclusive.

In digital communication, moderating online content is one of the biggest challenges, as it must strike a balance between freedom of expression and protection from harmful content. This so-called content moderation refers to the process of reviewing and evaluating user contributions and, if necessary, restricting, amplifying or deleting them. This is necessary because countless new texts, images and videos are uploaded to social platforms every second. To prevent problematic content from circulating unchecked, platforms need clear rules for dealing with it. These rules can come from laws, such as the provisions of the European Digital Services Act (DSA). In addition, platforms such as TikTok, YouTube and Instagram have their own community guidelines. These are documented in guidelines and FAQs and specify which content may and may not be shared by their users.

Why are rules needed for dealing with content on the internet?

Moderation usually focuses on content that contains hate speech, insults or deliberately spread disinformation. If these are not identified and dealt with in a timely manner, they can spread unchecked. In such cases, there is an increased risk that discriminatory or false information will gain widespread reach, distort public debate and undermine trust in platforms.

One example of this is the situation of the Rohingya in Myanmar, a Muslim minority that has been

systematically discriminated against and persecuted there for decades. In 2017, anti-Rohingya content was disseminated via Facebook, contributing to violence, human rights violations and displacement (Amnesty International, 2022). Another particularly drastic case is the murder of Ethiopian professor Meareg Amare in 2021. Hate speech and disinformation about him circulated repeatedly on Facebook and were not removed despite multiple reports. This fact later became the basis for a class action lawsuit against Facebook's parent company Meta, seeking damages for failure to moderate harmful content and thus complicity in the murder (Milmo, 2025).

Such events show that if dangerous content is not moderated in a timely manner, the consequences can be serious. This leads to a key question: who actually decides which posts remain online and which do not?

Who decides what stays online: humans or machines?

Content moderation is not a new phenomenon that only emerged with digital platforms. Harmful content, misinformation and insults have always existed, as have debates about what should be published. Newspapers have also always selected which texts or images appear, especially in sensitive areas such as war and violence. In Germany, for example, journalists follow the Press Code (German Press Council, 2025). This ensures source-based, trustworthy and responsible reporting. However, the internet has added a new dimension to this task. Today, content spreads around the globe in a matter of seconds and must be checked in real time to ensure that it does not violate national laws or the community standards of the respective platform. National laws and cultural contexts vary greatly from place to place and region to region. This makes moderation particularly challenging because a post that is unproblematic in one country may clearly violate applicable law or social norms in another.

In order to cope with the flood of information and moderation requests, machines and humans now share the work. Algorithms take on a large part of the preliminary work. They sort content, categorise it and automatically perform certain steps. Unclear or particularly sensitive cases are passed on to human moderators. In theory, this division of labour is intended to reduce the volume of work so that humans have more time for difficult decisions.

When moderation reaches its limits

In practice, however, this hope is not always fulfilled. Algorithms identify and categorise content such as text, images or videos by recognising patterns or comparing them with databases. However, they quickly reach their limits when it comes to understanding the meaning of language, symbols or context. A word that is considered harmless slang in one country may be seen as an insult in another. Similarly, images or gestures can carry very different meanings depending on their social or cultural environment.

One example is the 👌 emoji. In the United States and many other Western countries, it expresses approval, while in Brazil the same gesture is considered offensive – much like showing the middle finger in Germany (Moran, Abramson & Moran, 2014, p. 352). The boundaries between satire and insult, irony and criticism, or between legitimate expression and hate speech are often blurred.

However, this is not merely a technical problem, as even humans are not always able to interpret such nuances clearly. Moreover, (semi-)automated systems used in content moderation – meaning the combination of algorithmic pre-filtering and human decision-making – do not make neutral judgements. The following section explains why these systems often reinforce existing inequalities in the world instead of reducing them.

How does content moderation reinforce global inequalities?

Research shows that the cases that remain for human moderators are often the most complex and distressing ones (see Roberts, 2019). Particularly problematic is the fact that major platform companies frequently outsource this moderation work to countries of the so-called *Majority World*[1]. In places such as Kenya or the Philippines, people are required to review highly disturbing material under precarious working conditions. They often carry out the "last mile" of moderation in circumstances that exacerbate psychological trauma, economic insecurity, and invisibility (Gray/Suri, 2019). This work can therefore have severe consequences for their health (Dachwitz, 2024).

This distribution of labour reflects the colonial legacies of global economic inequality. It follows the logic of historical exploitation that continues to shape the present. While countries and communities in the Majority World still shoulder much of the low-paid and psychologically taxing work, the economic profit remains with large platform companies in the Global North (see Siapera, 2022).

However, it is not only the distribution of labour that reinforces inequalities; the process of content moderation itself also plays a role. The training data, categories, and norms by which algorithms filter, sort, and evaluate content are predominantly shaped in the Global North. As a result, cultural expressions are often misclassified or stigmatised (Noble, 2018; Eubanks, 2017). Marginalised voices are disproportionately deleted, while hate speech remains. For example, feminist or queer content is frequently deleted on the grounds of 'nudity' (Dalek et al., 2021), while misogynistic and sexist material continues to spread virally (Regehr et al., 2024).

Furthermore, platforms invest very little in moderating content in so-called minority languages, rendering communities in the Majority World increasingly invisible (De Gregorio & Stremlau, 2023). Content moderation is therefore far from a purely technical process. It is a deeply political field, that shapes relations of power and visibility, and that can either foster participation or intensify inequality.

Understanding and making inequalities visible

To reveal these structures, feminist, intersectional, and decolonial approaches are particularly useful. They shed light on who is most affected by these inequalities and how different forms of discrimination intersect. This includes women, gender-diverse people, racialised communities, and linguistic minorities. Precariously employed content moderators are also exposed to multiple layers of strain.

Intersectionality describes how various forms of discrimination – such as gender, race, or social class – overlap and reinforce one another. Building on this, decolonial perspectives call for a critical examination of traditional power relations and for centring the voices of those who have been marginalised.

So how can content moderation be designed in a way that does not reinforce existing power structures, but instead promotes participation and helps to reduce inequality?

How can intersectional and decolonial approaches be incorporated into content moderation?

These challenges cannot be solved by better technologies alone. Intersectionality should therefore be understood as a fundamental design principle of socio-technical systems — ensuring that marginalised groups have a seat at the table (Crenshaw, 1989; Gebru et al., 2021; D'Ignazi o & Klein, 2020).

Decolonial and feminist perspectives further call for linguistic and cultural diversity to be systematically considered in the development process, for local knowledge to be recognised, and for community-based models to be strengthened. A central element of this is participatory design — involving affected communities from the very beginning of technology development. Rather than building AI systems *about* marginalised groups, the goal is to design them *with* them.

Examples of such participatory approaches can be found in community-based projects developing data-based AI systems to detect and prevent gender-based violence. In the context of femicide prevention, for instance, where those affected act not only as data sources but as co-designers (D'Ignazio 2024; Suresh et al. 2022). These approaches require profound structural change: within the institutions that regulate AI, the companies that develop it, and the social processes that determine whose perspectives matter.

There are also initiatives that embed local languages and knowledge in datasets to preserve cultural diversity. The Data Justice Lab and the DAIR Institute founded by Timnit Gebru demonstrate how research and practice can work hand in hand to promote decolonial and gender-aware AI.

Especially in the context of (semi-)automated systems for content moderation, intersectional and decolonial approaches can help ensure that moderation practices no longer reproduce existing inequalities, but instead make diversity visible and enable participation. Only then can this field become more than a space for deletion and filtering, a space that fosters care, resistance, and reflection (Costanza-Chock, 2020; Benjamin, 2019).

Inside automation: An anthology on content moderation from the perspective of the majority world

The growing discussion around intersectional and decolonial approaches highlights the need for spaces where the marginalised voices of content moderators themselves can be heard — allowing their experiences of moderation work to inform debates about how to create fairer structures. Such a space is being developed within the Human in the Loop? research project.

The project's second case study explores how human–machine interaction in content moderation can be shaped effectively. The edited volume brings together theoretical analyses, empirical case studies, and first-hand accounts. Voices from the Majority World not only shed light on the everyday pressures and lived realities of content moderators within the global structures of the platform economy; they also highlight practices of resistance and approaches that are central to building more just digital infrastructures.

Without these perspectives, any discussion of content moderation remains incomplete — and any attempt at reform risks reinforcing existing power relations instead of transforming them.

The edited volume is scheduled for publication in summer 2026 with transcript Verlag.

Comments

[1] The term refers to regions where the majority of the world's population lives, such as Africa, parts of Asia, and Latin America. It is used deliberately to offer an alternative perspective to terms such as 'Global South' or 'developing countries', which are often associated with a deficit-oriented view.

References

Amnesty International (2022). Myanmar: Facebook-Algorithmen haben Gewalt gegen Rohingya befördert. https://www.amnesty.de/allgemein/pressemitteilung/myanmar-facebook-algorithmen-haben-gewalt-gegen-rohingya-befoerdert

Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity Press.

Costanza-Chock, S. (2020). Design justice: Community-led practices to build the worlds we need. The MIT Press.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine. University of Chicago Legal Forum, 1989(1), 139–167.

Dachwitz, I. (2024). Moderator:innen leiden unter schweren psychischen Erkrankungen. https://netzpolitik.org/2024/facebook-moderatorinnen-leiden-unter-schweren-psychischen-erkrankungen/

Dalek, J., Dumlao, N., Kenyon, M., Poetranto, I., Senft, A., Wesley, C., Filastò, A., Xynou, A. & Bishop A., (2021). No Access LGBTIQ Website Censorship in Six Countries. https://citizenlab.ca/2021/08/no-access-lgbtiq-website-censorship-in-six-countries/

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

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. The MIT Press.

Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Gray, M. L., & Suri, S. (2019). Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass. Houghton Mifflin Harcourt

Kloiber, J. (2023): Without Us, There Are No Social Media Platforms. https://superrr.net/en/blog/without-us-there-are-no-social-media-platforms

Lee, C., Gligorić, K., Kalluri, P.R., Harrington, M., Durmus, E., Sanchez, K.L., San, N., Tse, D., Zhao, X.,

Hamedani, M.G., Markus, H.R., Jurafsky, D. & Eberhardt, J.L. (2024). People who share encounters with racism are silenced online by humans and machines, but a guideline-reframing intervention holds promise, Proc. Natl. Acad. Sci. U.S.A. 121 (38) e2322764121, https://doi.org/10.1073/pnas.2322764121

Mignolo, W. D. (2011). The darker side of Western modernity: Global futures, decolonial options. Duke University Press

Milmo, D. (2025). Meta faces £1.8bn lawsuit over claims it inflamed violence in Ethiopia. https://www.theguardian.com/technology/2025/apr/03/meta-faces-18bn-lawsuit-over-claims-it-inflamed-violence-in-ethiopia

Moran, R.T., Abramson, N.R., & Moran, S.V. (2014). Managing Cultural Differences (9th ed.). Routledge. https://doi.org/10.4324/9781315871417

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.

Regehr, K., Shaughnessy, C., Zhao, M. & Shaughnessy, N. (2024). SAFER SCROLLING. How algorithms popularise and gamify online hate and misogyny for young people. https://www.ascl.org.uk/ASCL/media/ASCL/Help%20and%20advice/Inclusion/Safer-scrolling.pdf

Roberts, S. (2021). Behind the Screen: Content Moderation in the Shadows of Social Media. New Haven and London: Yale University Press

Siapera, E. (2022). Content Moderation, Racism and (de)Coloniality. Int Journal of Bullying Prevention 4, 55–65. https://doi.org/10.1007/s42380-021-00105-7

Suresh, H., Movva, R., Lee Dogan, A., Bhargava, R., Cruxen, I., Martinez Cuba, A., Taurino, G., So, W. & D'Ignazio, C. (2022). Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Feminicide Counterdata Collection. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 667–678. https://doi.org/10.1145/3531146.3533132