doi:10.1017/aju.2023.23

SYMPOSIUM ON CRITICAL INTERNATIONAL LAW AND TECHNOLOGY

BIAS IN SOCIAL MEDIA CONTENT MANAGEMENT: WHAT DO HUMAN RIGHTS HAVE TO DO WITH IT?*

*Dorothea Endres**, Luisa Hedler***, and Kebene Wodajo*****

In a global context where political campaigning, social movements, and public discourse increasingly take place online, questions regarding the regulation of speech by social media platforms become ever more relevant. Companies like Facebook moderate content posted by users on their platforms through a mixture of automated decision making and human moderators.[1] In this content moderation process, human rights play an ambiguous role: those who struggle with marginalization may find a space for expression and empowerment, or face exacerbation of pre-existing bias.[2] Focusing on the role of human rights in Meta's content management, this essay explores how the protection of speech on social media platforms disadvantages the cultural, social, and economic rights of marginalized communities. This is not to say that speech on social media platforms is devoid of emancipatory potential, but that this potential is not uniformly or equally accessible. We see the incorporation of human rights considerations into decision-making processes as an avenue for alleviating this challenge. This approach faces obstacles from the platforms' business models, which decenters human rights concerns, and from the limitations of liberal accounts of human rights. From within and against these constraints, human rights can be mobilized as emancipatory power in an effort to decrease marginalization.

Bias in Freedom of Speech

Norms protecting speech as a human right have developed and continue to be reiterated through practice dominated by the values and interests of those already privileged.[3] In fact, there is a general tendency to hierarchize and privilege dimensions of individual liberties, to minimize societal concerns, and to frame freedom of speech in a

* The authors would like to thank Abhimanyu George Jain and all reviewers for their helpful comments.

** *Research Associate for Legal Philosophy at the University of Geneva and PhD Candidate at the Graduate Institute, Geneva, Switzerland.*

*** *PhD Fellow at the Department of Business Humanities and Law of Copenhagen Business School, Copenhagen, Denmark.*

**** *Postdoctoral Fellow at the Institute for Business Ethics, University of St. Gallen, St. Gallen, Switzerland.*

[1] Facebook, *How Does Facebook Use Artificial Intelligence to Moderate Content?*

[2] Those who are privileged and disadvantaged vary within and across geographical spaces and context. *See* B.S. Chimni, *Third World Approaches to International Law: A Manifesto*, 8 Int'l Cmty. L. Rev. 3, 6–7 (2006); Ramesh Srinivasan, Whose Global Village? Rethinking How Technology Shapes Our World 53–54 (2017).

[3] *See generally* B.S. Chimni, International Law and World Order: A Critique of Contemporary Approaches 541 (2d ed. 2017).

manner that benefits the privileged.[4] The presence of this hierarchization does not preclude the use of human rights in struggles that do not align with those interests,[5] but leads to the incorporation of bias into the way in which freedom of speech can be exercised.[6]

There is a double standard for the conceptualization of what it means to exercise free speech. Those who are already privileged operate from a sphere of protection closely aligned with their practices.[7] For instance, the traditional dress of Bavarian women, as seen during Oktoberfest and involving sexualized displays of their breasts, does not interfere with their ability to engage in speech on Facebook.[8] In contrast, the Brazilian National Foundation of Indigenous Peoples had their Facebook account suspended in 2018 for a post about traditional knowledge portraying two Waimiri Atroari in their traditional dress, which leaves the nipple bare[9]—the nipple being the one body part Facebook explicitly bans—in order to prevent "uncivilized" pornographic material, inconsistent with Western values.[10] Against this bias toward Western values, the Brazilian Indigenous women had to defend their dress to the extent of needing a judicial decision to have their accounts reinstated—before being able to exercise their right to speak. In short, Facebook's restrictions on nudity differ across cultures with regard to their impact on speech.

This structure has substantial marginalizing effects. In Ratna Kapur's words, "the liberal tradition from which human rights have emerged not only incorporates arguments about freedom and equal worth . . . it also incorporates civilization, cultural backwardness, racial and religious superiority."[11] In our example, bare nipples do not accord with Facebook's standard of civilization, while expansive décolletage does. Consequently, Bavarian women are immediately free and "equal," while Indigenous women need to take additional steps "to climb up" into that "free" space: they are pushed further to the margin.[12]

Structural Bias in Social Media Content Management

Increasingly, social media platforms are not only spaces for human rights struggles to exercise free speech and document evidence of rights violations, but also spaces for the repression of rights. Popular social movements such as #BlackLivesMatter[13] demonstrate that marginalized groups can carve out a space on platforms to express their concerns. However, bias remains—often with exacerbating socioeconomic implications. For instance, Facebook uses seemingly neutral knowledge to enable gentrification: "Lookalike audience" allows one to advertise

[4] Tony Evans, *Castles in the Air: "Universal" Human Rights in the Global Political Economy*, *in* Relations of Global Power: Neoliberal Order and Disorder 152, 163 (Gary Teeple & Stephen McBride eds., 2011); Upendra Baxi, The Future of Human Rights 39 (2d ed. 2008).

[5] Baxi, *supra* note 4, at 46.

[6] Balakrishnan Rajagopal, International Law from Below: Development, Social Movements, and Third World Resistance 235 (2003); Makau Mutua, Human Rights Standards: Hegemony, Law, and Politics 16 (2016).

[7] Facebook, *Community Standards, Nudity (2023).*

[8] See for a fairly sexualized version of that dress the Facebook page of "Dirndlkalender."

[9] Sérgio Matsura, *Facebook bloqueia conta da Funai por imagem de indígenas com seios nus*, O GLOBO (Aug. 30, 2018).

[10] The Oversight Board (OB) has highlighted similar concerns with regard to transgender individuals' bare breasts: OB, Gender Identity and Nudity, 2022-009-IG-UA and 2022-010-IG-UA.

[11] Ratna Kapur, *Human Rights in the 21st Century: Take a Walk on the Dark Side*, 28 Sydney L. Rev. 665, 674 (2006).

[12] The OB has denounced Facebook's insufficient consideration of marginalized groups with respect to a post in Arabic that aimed at reclaiming speech experienced as hurtful by LGBTQ+ communities. OB, Reclaiming Arabic Words, 2022-003-IG-UA.

[13] Black Lives Matter, #BlackLivesMatter (2013).

https://doi.org/10.1017/aju.2023.23 Published online by Cambridge University Press

housing to those who share similar interests.[14] Thus, empowerment arising out of #BlackLivesMatter may lead to increased barriers in the housing market in regions where there is less enthusiasm about that movement.

Prejudice against marginalized groups' content is particularly attributable to context-blindness that replicates existing structural inequalities. Content management relies on static rules that are not sufficiently attentive to meaning-defining contexts. This context-blindness regularly requires those at the margins to defend and explain how their specific context changes the meaning of their speech-act. In the Wampum belt case,[15] the Facebook Oversight Board (OB) assessed the removal of a post made by Indigenous North American artists in Canada for featuring the phrase "Kill the Indian/Save the Man" as a violation of Facebook's Hate Speech Community Standard. The post was intended to denounce the history of Native Americans having to renounce their identity in order to survive, as the applicant successfully explained to the OB, leading to readmittance of the post. The initial platform decision was blind to this specific context—Meta was unable to explain how two human reviewers failed to understand the phrase "the sole purpose is to bring awareness to this horrific story" so as to qualify the post as permitted counter-speech.[16] Despite explicit labeling of the post, the artists' speech was misaligned with presumably "neutral" but context-insensitive platform policy, with the result that they had to carve out their space for political activity before being able to use their "freed" speech. In contrast, the display of a weapon, a statement that is arguably as supportive of violence as the word "kill," is not automatically removed—sometimes even after the authors have been banned for illegal trade in weapons.[17]

Bias in platforms' content management tends to replicate and amplify existing vulnerabilities and inequalities with implications not only for free speech but also for socioeconomic rights [18] For instance, human rights groups criticized Facebook and Twitter for systematically silencing protests about the livelihood-disrupting conditions of Palestinians in Sheikh Jarrah.[19] Such asymmetries in take-downs have been highlighted by human rights activists more generally, to the extent that the advised strategy for investigations on human rights violations is to race against the take-down time to save the social media content off-line.[20] This adds another burden in the struggle to denounce any human rights violation.

The incorporation of human rights considerations into a platform's architecture is a possible avenue for alleviating the potential biases of decision making in social media.[21] However, there are significant obstacles for such emancipatory use, as will be detailed below.

[14] Rebecca Kelly Slaughter, Janice Kopec & Mohamad Batal, *Algorithms and Economic Justice: A Taxonomy of Harms and a Path Forward for the Federal Trade Commission*, 23 YALE J. L. & TECH. 1, 19–21 (2021).

[15] OB, Wampum Belt, 2021-012-FB-UA (2021).

[16] *Id.* The understandable reason to protect survivors from pain the art could cause was only acknowledged by the artist—not by Meta. Mastodon for instance provides for a technological solution, "trigger warning," that aims to provide protection in such cases. Jacqueline Burggraf, Laura Gil & Anna Kuschezi, *Content Warnings on Mastodon*, H_DA (Jan. 18, 2023).

[17] Elizabeth Dwoskin & Naomi Nix, *Facebook's Ban on Gun Sales Gives Sellers 10 Strikes Before Booting Them*, WASH. POST (June 9, 2022).

[18] VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018).

[19] Human Rights Watch, *Israel/Palestine: Facebook Censors Discussion of Rights Issues* (2021); Access Now, *Sheikh Jarrah: Facebook and Twitter Systematically Silencing Protests, Deleting Evidence* (updated Jan. 26, 2023).

[20] Anna Veronica Banchik, *Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content*, 23 NEW MEDIA & SOC'Y 1527, 1535–38 (2021).

[21] Nicolas Suzor et al., *Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online: Gender-Based Violence Online*, 11 POL'Y & INTERNET 84 (2019).

https://doi.org/10.1017/aju.2023.23 Published online by Cambridge University Press

Human Rights Are Not the Main Priority

Most social media platforms' policies and rules, particularly that of Facebook and its OB, explicitly refer to the human rights responsibilities of platforms in speech regulation. Nevertheless, the application of human rights standards remains secondary to Meta's articulated values. Those values focus—far from proclamations of liberty and equality—first and foremost on "giving voice," not on ensuring equal and free speech.[22] This is well aligned with the corporation's business model: "more speech" means more profit.[23] In this setting, the protection of (other) human rights is regularly secondary to the corporation's priorities.[24] Assessing the Myanmar violence in 2017, Amnesty International pointed to Facebook's business model—geared toward leaving content online, even if it fuels hatred—as a major reason for the belated and inadequate content review that allowed for a vast range of human rights violations.[25] Leaked internal documents demonstrate that Facebook knew that a subsequent change in its algorithm in 2018 increased divisive and violent speech leading to human rights violations.[26] So, instead of reassessing its structure in order to reflect human rights, Facebook exacerbated what had caused human rights violations before, knowing, as said files testify as well, that content-moderation was unable to keep hate speech in check.[27]

In this regard, the OB's role in highlighting ambiguities in Meta's rules and policies in consideration of human rights provisions could be viewed in a positive light,[28] but its lack of diversity, limited impact, and financial dependence on Meta limit this optimism.[29] In fact, that the OB reviewed 176 recommendations out of over 2.5 million cases qualifies not even as a drop in the ocean.[30] In sum, Facebook's structure is less attuned to a fairly balanced marketplace of ideas than to a marketplace for revenue creation through advertisement.[31]

Limitations in Human Rights

To the extent that human rights considerations are implicated in the framework for decisions on free speech in social media, the process has to contend with the bias implicit in human rights. That is, there is a tendency to hierarchize and privilege dimensions of individual liberties with minimal consideration to societal dimensions that cannot always be addressed through protections for individual liberty.[32] In fact, the focus on speech exacerbates the tendency to overemphasize individual liberty, to the detriment of considerations about social, economic, and cultural rights. In the example of the Indigenous women in Brazil,

[22] Meta, *Our Culture* (2023).

[23] Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019).

[24] Sanna Spišák, Elina Pirjgtanniemi, Tommi Paalanen, Susanna Paasonen & Maria Vihlman, *Social Networking Sites' Gag Order: Commercial Content Moderation's Adverse Implications for Fundamental Sexual Rights and Wellbeing*, Social Media 1 (2021).

[25] Amnesty International, *Myanmar: The Social Atrocity: Meta and the Right to Remedy for the Rohingya* (Sept. 29, 2022).

[26] Keach Hagey & Jeff Horwitz, *The Facebook Files, Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead*, Wall St. J. (Sept. 15, 2021).

[27] Deepa Seetharaman, Jeff Horwitz & Justin Scheck, *Facebook Files, Facebook Says AI Will Clean Up the Platform. Its Own Engineers Have Doubts*, Wall St. J. (Oct. 17, 2021).

[28] David Wong & Luciano Floridi, *Meta's Oversight Board: A Review and Critical Assessment*, Minds & Machines (2022).

[29] Lakshmi Gopal, *Facebook's Oversight Board & the Rule of Law: The Importance of Being Earnest*, ABA (2021).

[30] Facebook, *Quarterly Transparency Report*, 2–3 (4th quarter 2022).

[31] Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads (2016).

[32] Danielle K. Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 Boston U. L. Rev. 1435 (2011).

https://doi.org/10.1017/aju.2023.23 Published online by Cambridge University Press

Facebook's community guidelines do not provide for a group right to culture, but require that the post is taken down—that freedom of speech is reclaimed and reassessed in light of the interest *to talk* about cultural rights to dress in traditional clothing.[33]

Furthermore, most biases on platforms are manifestations of existing socioeconomic hierarchies and marginalization that cannot be addressed through the individualistic lens of human rights.[34] The marginalization of Indigenous communities, amplification of disinformation against minorities in Myanmar, and silencing of Palestinians are all rooted in pre-existing inequalities. The root of such marginalization is more profound than an incident of individual rights violation,[35] and the resultant harm, such as perceived and reported discrimination, is a symptom of a deeper structural problem such as racism and other forms of prejudice.[36]

It is, thus, crucial to pay attention to potential obstacles such as limitations within the human rights system itself and platforms' business models, which decenter human rights concerns. With those priorities in mind, the effort to incorporate human rights considerations into platforms' system of speech regulation may become a positive step toward countering bias.

Human Rights to Counter Bias

Structural bias in human rights and content management need not be mutually reinforcing. In fact, the emancipatory potential of human rights can play a part in resistance against prejudice in and beyond platforms' content management. There is a conceivable space for human rights work and struggle on social media platforms. The transnational and accessible character of social media networks is enabling global cooperation around the exposure of human rights violations by counter-power movements.[37] For instance, the movements during the Arab Spring relied on communication via Facebook.[38] While this appraisal has been celebrated as a Facebook revolution, movements less aligned with Western values received less favorable treatment. In other words, as long as the social movement promotes what the West deems "universal," social media platforms provide a welcoming space to promote their cause.[39]

For those adversely affected, in order to overcome the limits of human rights in addressing structural bias on platforms, a focus on human rights' emancipatory power is promising. Human rights claim-making can push beyond the structural constraints of rights and can organize and unify people in their struggle against bias.[40] Returning to our example of the Brazilian Indigenous women, their cultural rights provided them with a tool to combat exclusion. For Facebook, a next step toward equality and inclusion would mean, on the one hand, to push these cultural rights further into the structure of Facebook's content management so that takedown protocols are more careful in the assessment of cultural rights, and on the other hand, to expand and

[33] Meta, *Adult Nudity and Sexual Activity*.

[34] *See* Anja Bechmann, *Data as Humans: Representation, Accountability, and Equality in Big Data*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 73 (Rikke Frank Jørgensen ed., 2019).

[35] Kebene Wodajo, *Mapping (In)visibility and Structural Injustice in the Digital Space*, 9 J. RESPONSIBLE TECH. 100024 (2022).

[36] SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (1st ed. 2019).

[37] Rebecca J. Hamilton, *Governing the Global Public Square*, 62 HARV. INT'L L.J. (2021).

[38] Sara Reardon, *Was the Arab Spring Really a Facebook Revolution?*, NEW SCIENTIST (2012).

[39] *Id.*

[40] ISSA G. SHIVJI, THE CONCEPT OF HUMAN RIGHTS IN AFRICA (1989).

https://doi.org/10.1017/aju.2023.23 Published online by Cambridge University Press

protect the digital space in which Indigenous communities can discuss, promote, and transform their cultural rights.[41]

Conclusion

Free speech is a central object of concern both in a human rights framework and in the regulation of content on social media platforms. It can become a site of contention when hegemonic values encounter positions and inputs from those at the margins challenging said values. Certain kinds of speech are more robustly protected than others, thereby marginalizing those whose practices are at odds with dominant speech acts. In other words, the way in which freedom of speech is protected, both as a human right and in social media, is biased. While there are emancipatory, unifying elements in a human rights framework that carry the potential for using it as a framework to better regulate social media, without careful consideration there is a risk of merely reinforcing and amplifying bias when selectively protecting free speech. Most importantly, typical human rights formulations are unable to capture structural bias, and therefore tend to neglect or even exacerbate such problems. From within and against those constraints, these platforms' potential as a space for the struggles of human rights movements—mobilizing the emancipatory power of human rights—may provide a path to counter bias, and toward positive transformations of rights and social media platforms for marginalized peoples.

[41] See for an example of such space creation: Godfried Asante, *"Where Is Home?" Negotiating Comm(unity) and Un/Belonging Among Queer African Migrants on Facebook*, *in* Queer and Trans African Mobilities: Migration, Asylum and Diaspora 135 (B. Camminga & John Marnell eds., 2022).

https://doi.org/10.1017/aju.2023.23 Published online by Cambridge University Press