Governing hate speech detection on online platforms: A human rights approach

Eva Nave[1] and Lottie Lane[2]

Abstract

The Internet is a global forum largely governed by private actors driven by profit concerns, often disregarding the human rights of historically marginalised communities. Increased attention is being paid to the corporate human rights due diligence (HRDD) responsibilities applicable to online platforms countering illegal online content, such as hate speech. At the European Union (EU) level, cross-sector initiatives regulate the rights of marginalised groups and establish HRDD responsibilities for online platforms to expeditiously identify, prevent, mitigate, remedy and remove online hate speech. These initiatives include the Digital Services Act, the Audiovisual Media Services Directive, the Directive on Corporate Sustainability Due Diligence, the Artificial Intelligence Act and the Code of conduct on countering illegal hate speech online. Nevertheless, the HRDD framework applicable to online hate speech has focused mostly on the platforms' responsibilities throughout the course of their operations - guidance regarding HRDD requirements concerning the regulation of hate speech in the platforms' Terms of Service (ToS) is missing. This paper[3] critically employs a conceptualisation of criminal hate speech as explained in the Council of Europe Committee of Ministers' Recommendation CM/Rec(2022)16, Paragraph 11, to develop specific HRDD responsibilities. We argue that online platforms should, as part of emerging preventive HRDD responsibilities within Europe, respect the rights of historically oppressed communities by aligning their ToS with the conceptualisation of criminal hate speech in European human rights standards.

1 Introduction

This paper addresses a vacuum in the legal framework by clarifying corporate human rights responsibilities in Europe to counter the most serious forms of online hate speech. In particular, we examine (emerging) standards on human rights due diligence (HRDD), artificial intelligence and online content moderation at the international and European level. We claim that there is a legal standard emanating from the HRDD framework in the European context prescribing the responsibility for online platforms - particularly for very large online platforms, video-sharing platforms and for platforms under the scope of the Directive on corporate sustainability due diligence[4] - to align their terms of service with the conceptualization of the criminal hate speech in Paragraph 11 of the Council of Europe Committee of Ministers Recommendation CM/Rec(2022)16.[5] Based on this claim, we provide recommendations to law- and policy-makers.

Around two thirds of the world's population are active Internet users.[6] While the Internet enables individuals to access information and exercise their freedom of expression, it also enables the proliferation of online hate speech. 'Online hate speech' broadly refers to discriminatory expressions shared on digital environments targeting historically or systematically marginalized[7] people. Recommendation CM/Rec(2022)16 reiterates that hateful expressions represent a violation of human rights. When unaddressed, these can hinder peace and development by denying the values of pluralism, tolerance and broadmindedness essential in a democratic society.

The rise of online hate speech results from specific features of the Internet. First, unlike in traditional media, most con-

[1] PhD candidate, Center for Law and Digital Technologies, Leiden University, https://orcid.org/0000-0001-5040-1601, e.v.r.nave@law.leidenuniv

[2] Assistant Professor of Public International Law, University of Groningen, https://orcid.org/0000-0002-2986-3075, c.l.lane@rug.nl

[3] The completed version of this paper is published in the Computer Law and Security Review, Special Issue on Vulnerability, Marginalization and Data Protection Law, available at: https://doi.org/10.1016/j.clsr.2023.105884

[4] European Commission (2022) Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937.

[5] Council of Europe Committee of Ministers, Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech (CM/Rec(2022)16).

[6] Number of internet and social media users worldwide as of January 2023 (2023) <https://www.statista.com/statistics/617136/digital-population-worldwide/> accessed 26 April 2024.

[7] This paper uses 'oppression' and 'marginalisation' interchangeably.

tent published on the Internet can be quickly shared with little to no monitoring, made available to large audiences, published under anonymity, and easily manipulated in ways that intensify hate (e.g. hate profiles, memes and deep fakes). Second, online content is hosted by businesses primarily driven by profit goals, often at the expense of human rights. The potentially negative impact of AI-driven content moderation by online platforms is under increasing scrutiny. For example, Meta Platforms, Inc. (formerly named Facebook, Inc.) faces legal action for alleged negligence in facilitating the genocide of Rohingya Muslims in Myanmar after its algorithm failed to remove hateful posts and amplified hate speech.[8] Similarly, whistle-blower Frances Haugen alerted that Facebook neglected reports of accounts and hate speech content towards Muslims in India, potentially leading to offline violence.[9] There are reportedly other situations of human rights abuses by different platforms.[10]

2 Responses by scholars, legislators and policy-makers

Legal scholars have emphasized the growing impact of social media platforms in the application of regulatory frameworks for freedom of expression and democratic processes and the subsequent need to expand the legal scholarship focusing on the regulation of online platforms.[11] In this context, it is relevant to consider that most online platforms are based in the USA and thus typically bound by the USA framework on freedom of expression, corporate human rights due diligence and intermediary liability. Conversely, to the extent that online platforms operate in European Union (EU) territory, they must also abide by the regional human rights frameworks in Europe, which differ significantly from those applicable in the USA. The reconciliation of different regional standards has been challenging, not only for online platforms but also for judicial bodies in enforcing their decisions.[12]

Legislators and policy-makers at the international, regional and national level have made efforts to prevent and address the negative impact of business on human rights, including through HRDD and through liability regimes.[13] The HRDD regime includes the seminal United Nations Guiding Principles on Business and Human Rights (UNGPs), which are arguably the most authoritative international expression of the corporate responsibility to respect human rights through

[8] Al Jazeera, 'Rohingya sue Facebook for $150bn for fuelling Myanmar hate speech' (7 December 2021) <https://www.aljazeera.com/news/2021/12/7/rohingya-sue-facebook-for-150bn-for-fuelling-myanmar-hate-speech> accessed 26 April 2024.

[9] Al Jazeera, 'Facebook failing to check hate speech, fake news in India: Report' (25 October 2021) <https://www.aljazeera.com/news/2021/10/25/facebook-india-hate-speech-misinformation-muslims-social-media> accessed 26 April 2024.

[10] Shaun Harper, 'Hate Speech Rises On Twitter After Elon Musk Takes Over, Researchers Find' (*Forbes*, 31 October 2022) <https://www.forbes.com/sites/shaunharper/2022/10/31/elon-musk-twitter-takeover-leads-to-n-word-and-hate-speech-increase-lebron-james-calls-for-action/?sh=f28a381dd99a> accessed 6 April 2023; Hadi Al Khatib and Dia Kayyali, 'YouTube Is Erasing History' (*The New York Times*, 23 October 2019) <https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html> accessed 26 April 2024.

[11] E.g. Kate Klonick, 'The new governors: The people, rules, and processes governing online speech'(2017) Harv. L. Rev., 131, 1598; Tarlach McGonagle, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation' (2020) Oxford Handbooks in Law (pp. 467–485), 10; Tarlach McGonagle, 'The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing', 232, in Rikke Frank Jørgensen (eds), 'Human Rights in the Age of Platforms' (2019) Cambridge, MA: The MIT Press <https://doi.org/10.7551/mitpress/11304.001.0001> accessed 26 April 2024; Judit Bayer, Bernd Holznagel, Päivi Korpisaari (ex. Tiilikka), Lorna Woods, Volume 1' (2021) Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG., 30, <https://doi.org/10.5771/9783748929789> accessed 26 April 2024; Martin Moore and Tambini Damian (eds), 'Regulating Big Tech: Policy Responses to Digital Dominance' (2021), <https://doi.org/10.1093/oso/9780197616093.001.0001> accessed 26 April 2024.

[12] E.g. *Ligue contre le racisme et l'antisémitisme et Union des étudiants juifs de France c. Yahoo! Inc. et Société Yahoo! France (LICRA v. Yahoo!)*; and *Yahoo Inc. v LICRA*; European Court of Justice, Opinion of Advocate General Szpunar delivered on 8 June 2023 (1) Case C-376/22 clarifies that Union law prescribes the possibility for Member States to restrict the freedom to provide information society services to 'fight against any incitement to hatred on grounds of race, sex, religion or nationality, and violations of human dignity concerning individual persons'.

[13] Council of Europe, Committee on Artificial Intelligence, Consolidated Working Draft of the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, CAI(2023)18 < https://rm.coe.int/cai-2023-18-consolidated-working-draft-framework-convention/1680abde66> accessed 26 April 2024. Though State-centric and though not including standards directly applicable to companies, this Framework Convention provides key guidance for States regarding human rights centred approached to the governance and regulation of artificial intelligence.

HRDD.[14] At the EU level, the Directive on corporate sustainability due diligence (CSDDD) was just recently adopted.[15] Businesses - including online platforms - falling under the scope of the proposal should identify, prevent, mitigate and bring an end to negative impacts on human rights. Furthermore, the EU Artificial Intelligence Act (AI Act) emphasizes the need for protection of human rights in the digital environment.[16]

Concerning HRDD and moderation of harmful content online, in November 2022 the Regulation for a Digital Services Act entered into force.[17] The Digital Services Act adds to the EU Audiovisual Media Services Directive (AVMSD)[18] and enhances cross-sector due diligence responsibilities for digital services to remove illegal content online. This includes hate speech.[19] The due diligence framework in the Digital Services Act aligns with CM/Rec(2022)16 and builds on the Code of conduct on countering illegal hate speech online whereby IT companies commit to expeditiously review and remove hate speech and to promote transparency towards users.[20]

3 Gaps in law and policy

Despite these advancements, the HRDD framework applicable to online hate speech has focused mostly on explaining the responsibilities of companies throughout their operations. Guidance regarding HRDD requirements for the regulation of hate speech in the terms of service is missing. A key aspect remains unaddressed: how online platforms should define hate speech and how this should be communicated to their users. More specifically, is there a legal standard emanating from the European HRDD framework prescribing the responsibility for online platforms[21] to align their terms of service, as a minimum legal standard, with the conceptualization of the criminal hate speech as explained in the European human rights standards, in particular with the Recommendation CM/Rec(2022)16?

4 Scope of research

To answer this research question, we employ doctrinal research to differentiate between hate speech that is criminally actionable and hate speech that should be prohibited under civil or administrative law (Section 2). We focus on hate speech that is criminally actionable. The limitation of the requirement to harmonize and reflect the conceptualization of *criminal* hate speech is justified by a growing human rights understanding of criminal hate speech as reflected in CM/Rec/(2022)16, Para. 11, from which specific HRDD responsibilities can be developed. Building upon critical race, black feminist and intersectionality theories, this paper claims that platforms should explicitly conceptualize hate speech as discriminatory communications that target an open-ended list of historically or systematically oppressed people or group of people. This conceptualization should also consider the intersectionality of systems of oppression as an aggravating harm resulting from hate speech. It is worth remembering that the European Commission proposed to add hate speech to the list of EU crimes which, if and when this proposal materializes, will strengthen the need for a standardized conceptualization of criminal hate speech in online platforms' terms of service. This legal avenue supports compliance with the transparency and clarity required on Terms and Conditions (Article 14 Digital Services Act) generally imposed on all providers of intermediary services.

5 Key findings

In Sections 3 and 4, we investigate the HRDD regime.[22] We examine the HRDD framework applicable to AI businesses

[14] UN Human Rights Council, 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie' (2011) A/HRC/17/31. We use the term 'responsibility' to denote non-legally binding standards and 'obligation' when discussing binding standards.

[15] European Commission (2022) Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937.

[16] European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts COM(2021) 206 final, Explanatory Memorandum, 1.1.

[17] European Union, Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, Article 93.

[18] Directive 2010/13/EU of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), OJ L 95.

[19] European Commission (2018) Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, OJ L 63/50.

[20] European Commission (2016) The Code of Conduct on countering illegal hate speech online.

[21] 'AI businesses' is sometimes used synonymously with 'IT companies' or by 'internet intermediaries' (or 'intermediaries'), depending on the legal instrument under analysis.

[22] As a regulatory approach distinct from that of HRDD – as seen in the separate chapters on each regime in the Digital Services Act –, the EU liability regime for internet service providers (ISP) falls outside the remit of this research. These regimes are nevertheless related in that liability may follow from non-compliance with HRDD responsibilities. For discussion of ISP liability regimes and recent case law, see e.g. Andrea Bertolini et al., 'Liability of Online Platform: Study for the European Parliament' (2021) European Parliamentary Research Service PE 656.318; Berrak Genç-Gelgeç,

(the UNGPs, initiatives of the Organization for Economic Cooperation and Development, the CSDDD and the AI Act) and preventive HRDD responsibilities in moderation of illegal content, such as criminal hate speech (with reference to the Digital Services Act, the AVMSD, the Code of Conduct[23] and CM/Rec(2022)16). On this basis, we argue that terms of service fulfil the role of the human rights 'policy commitment' standard found in the UNGPs and should include a clear explanation of the platforms' commitment to human rights, including the prohibition of criminal hate speech. The drafting and updating of the terms of service should be a means for online platforms to respond to the systemic risk of online hate speech and terms of service should explicitly reflect the HRDD responsibilities to prohibit, remove and report criminal hate speech to relevant public authorities.

We suggest that to improve legal coherence in countering online hate speech in the European context, online platforms should follow CM/Rec(2022)16 Para.11's conceptualization of criminal hate speech in their terms of service. Paragraph 11 builds upon binding and non-binding international human rights standards, such as the International Convention on the Elimination of All Forms of Racial Discrimination,[24] the Convention on the Prevention and Punishment of the Crime of Genocide,[25] International Covenant on Civil and Political Rights,[26] Article 20(2), etc. As a result, some of these elements can be claimed to broadly represent international human rights standards and could therefore be extrapolated to international preventive HRDD responsibilities to counter criminal hate speech.

6 Assessment of practice

Section 5 presents an empirical qualitative analysis of three case studies: Facebook, Twitter and YouTube. We assess the compliance of the platforms' terms of service with the European Court of Human Rights jurisprudence on criminal hate speech, and with the conceptualization of criminal hate speech in CM/Rec(2022)16. The platforms were selected because they: (1) fall under the scope of CSDDD; (2) are signatories to the EU Code of Conduct; and (2) qualify as very large online platforms (VLOPs) as defined in the Digital Services Act.[27] The three case studies demonstrate that although Facebook, Twitter and YouTube have each to a certain degree adopted terms of service prohibiting hate speech, none of them currently conceptualizes hate speech in a way that is consistent with human rights standards. More specifically, none recognizes the difference between hate speech criminally actionable and hate speech which may be prohibited under civil or administrative law. Moreover, none recognizes the specific HRDD responsibilities associated with countering criminal hate speech. The three case studies reveal the lack of alignment of content moderation practices by online platforms with the HRDD responsibilities to identify, mitigate, cease, remedy and inform about potentially adverse impacts on human rights.

7 Recommendations and conclusion

In summary, addressing law- and policy-makers, we recommend that the European Commission issues a best practice guideline (under Article 35(3) Digital Services Act and Article 13 CSDDD) suggesting that VLOPs, and particularly video-sharing platforms, should explicitly mention in their terms of service that they prohibit, remove and report to law enforcement authorities criminal hate speech in line with the conceptualization in Paragraph 11 CM/Rec(2022)16. Further to this and also by issuing a best practice guideline, we recommend that the European Commission suggests that VLOPs, with a similar heightened focus on video-sharing platforms, adopt HRDD compliant content moderation processes which should likewise be explicitly mentioned in their terms of service.

This paper has primarily addressed the first phase of HRDD processes, i.e. the adoption of a policy commitment as a preventive HRDD responsibility. Further research is necessary to examine what could be required in relation to the remaining phases of HRDD, i.e. the tracking and communicating implementation and results as well as the provision of remedies when applicable. For example, what online platforms moderating content should do to identify and prevent the promotion of criminal hate speech, and how they could effectively respond to these risks, should be the subject of further study.

'Regulating Digital Platforms: Will the DSA Correct Its Predecessor's Deficiencies?' (2022) 18 Croatian Yearbook of European Law and Policy 25; United States Supreme Court, Twitter v. Taamneh 598 US (2023).

[23] Some EU instruments use the problematic expression 'illegal hate speech', which could lead the reader to understand that there is legal hate speech. This is not the case. Hate speech is always illegal but it can be criminalised only in its most serious forms. For legal coherence purposes, this paper will refrain from using 'illegal hate speech' unless referring to the title of an instrument.

[24] UN General Assembly, International Convention on the Elimination of All Forms of Racial Discrimination, 21 December 1965, United Nations, Treaty Series, vol. 660.

[25] UN General Assembly, Convention on the Prevention and Punishment of the Crime of Genocide, 9 December 1948, United Nations, Treaty Series, vol. 78.

[26] UN General Assembly, International Covenant on Civil and Political Rights, 16 December 1966, United Nations, Treaty Series, vol. 999.

[27] i.e. they have 45 million or more average monthly active recipients of their service in the Union: Digital Services Act, Recital 76.