Countering online hate speech: how to adequately protect fundamental rights?

Nave, E.V.R.

Summary

This manuscript explores legal approaches compliant with human rights to counter hate speech on online platforms. There is a growing prevalence of hate speech on online platforms. Online platforms have developed self-regulatory polices to counter hate speech. However, such private regulatory frameworks often remain opaque, and lack democratic enforcement and remedy mechanisms. This research builds on a critical conceptualization of online hate speech deriving from the European regulatory and policy framework, to investigate and propose legal avenues for the strengthening of the human rights due diligence (HRDD) responsibilities of online platforms to counter criminal online hate speech whilst upholding fundamental rights.

Chapter 1 lays the groundwork by presenting the context and social relevance, by introducing the problem statement and research questions, and by explaining the methodology and the scope. Social media online platforms can be broadly described as Internet hosting services that store, review, promote, or demote user-generated content to the general public through groups, tailored newsfeed, and messaging applications. With more than half of the world's population as active users of social media online platforms and using these online environments to exercise basic human rights such as freedom of expression, freedom of assembly and association, the content disseminated and moderation policies implemented are increasingly impactful on a global scale. Human rights activists, whistleblowers former employees of online platforms, and even the United Nations, have warned that business models adopted by some social media companies have not only failed to take down but even amplified online hate speech. Further studies also alert to the fact that the increased hate speech in digital environments can result in offline hate speech and hate crime.

In reaction to these events and due to pressure by States, human rights activists and civil society, some online platforms started to self-regulate hate speech, to share data about the hate speech prevalence, and to create oversight boards for appeal procedures on content moderation. However, such self-regulatory efforts are often criticized for not aligning with human rights standards. Some of the main points criticized for not aligning with human rights relate to: i) the conceptualization of hate speech applied by platforms; ii) the mechanisms of enforcement for content moderation policies; iii) the remedies available for users to appeal content moderation decisions. In an effort to regulate and to democratically oversee the regulatory frameworks

established by businesses to counter online hate speech, both States and international and regional organizations have been producing sector-specific legal and policy instruments. Nevertheless, discussions have arisen regarding the effectiveness and adequacy of such regulatory frameworks in promoting the respect for the human rights of people targeted by hate speech.

Accordingly, this manuscript asks: Building on a critical conceptualization of online hate speech and, more specifically, on criminal hate speech, deriving from the European regulatory and policy framework, how can European legislators, both at the European Union and at the Council of Europe levels, clarify the human rights due diligence responsibilities of online platforms to counter online hate speech whilst upholding fundamental rights? The methodology employed is three-fold: doctrinal; comparative; and, interdisciplinary. Doctrinal legal research of legal and policy frameworks applicable to online hate speech seeks to clarify existing legal standards, loopholes, and potential future legal avenues. Comparative legal analysis is utilized to investigate the alignment, or lack thereof, between the standards adopted by online platforms through their terms of service and European human rights standards. Interdisciplinary research combines findings from legal, sociology, and digital technologies studies to systematize concerns and propose content moderation practices suitable to countering online hate speech and compliant with human rights.

Chapter 2 proposes a new legal conceptualization of hate speech in the European context. Current frameworks lack a standardized approach to the conceptualization of hate speech. Some conceptualizations are overbroad, and others are underinclusive; overbroad because they lead to the removal of legal content (e.g. removal tools deleting legal content posted by marginalized communities), and underinclusive as the context of posts by linguistic minorities is often disregarded. This Chapter analyses the European regulatory framework through the lens of the first legal conceptualizations of hate speech deriving from critical (race) theory and (black) feminist intersectionality theory. There are two main findings from this Chapter. First, this Chapter suggests that the European regulatory framework needs to explicitly acknowledge the conceptualization of hate speech by critical legal scholars as expressions intended to perpetuate historical or systematic oppression. Second, this Chapter advocates that the conceptualization of hate speech in the European context can only achieve legal cohesion when all European regulatory instruments expressly account for the intersectionality of systems of oppression.

Chapter 3 advances specific preventive HRDD responsibilities applicable to online platforms countering online hate speech. Increased attention is being paid to the corporate HRDD responsibilities applicable to online platforms to counter online hate speech. At the European Union 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. 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 is missing. This Chapter 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. This research includes an empirical qualitative analysis of three case studies: Facebook (Meta Platforms, Inc.), X Corp. (previously Twitter, Inc.), and YouTube. This empirical analysis assesses the compliance of the platforms' Terms of Service with the conceptualisation of criminal hate speech in CM/Rec(2022)16. This Chapter claims that online platforms should, as part of emerging preventive HRDD responsibilities within Europe, respect the rights of historically oppressed communities by aligning their Terms of Service with the conceptualisation of criminal hate speech in European human rights standards.

Chapter 4 proposes a new legal minimum standard expanding corporate human rights responsibilities of online platforms providing E2EE services to mitigate a category of criminal hate speech – incitement to violence. Services adopted by online platforms have enabled the proliferation of online hate speech. In particular, end-to-end encrypted (E2EE) services have been under increased scrutiny for hosting hate mongers. Legal practitioners and law enforcement struggle to conceptualise the responsibilities of E2EE services to not host hate speech without disproportionately affecting the users' rights to freedom of expression, association, privacy, or data protection. After establishing the general HRDD framework for Artificial Intelligence businesses corporate HRDD to mitigate criminal hate speech, this Chapter delves deeper into the digital technologies and encryption features used for content moderation in E2EE services. This analysis applies the HRDD framework coupled with homomorphic encryption, metadata, and hashing to selected criminal hate speech inciting to violence. Additionally, this Chapter clarifies the standards for cooperation between online platform and law enforcement in the context of incitement to violence in large group chats on E2EE services. To conclude, Chapter 4 proposes a new legal standard expanding corporate HRDD of online platforms providing E2EE services through the regulation and application of metadata, hashing, and homomorphic encryption to disrupt incitement to violence in large groups on E2EE services.

Chapter 5 proposes a comprehensive remedial responsibilities framework for online platforms which caused or contributed to criminal hate speech based on the general corporate human rights responsibilities framework. Legislators have developed binding legal frameworks clarifying the human rights due diligence and liability regimes of these platforms to identify and prevent hate speech. However, these legal frameworks fail to clarify the remedial responsibilities of online platforms to redress people harmed by criminal hate speech caused or contributed to by the platforms. Meta's contribution to the genocide

of the Rohingya in Myanmar is analysed as one of the most thoroughly documented cases showing the societal impact of the corporate human rights responsibilities of very large online platforms contributing to the amplification of criminal hate speech.

This Chapter investigates the application of the right to an effective remedy to cases of online hate speech. This investigation also examines the international standards on the right to remedy for cases of gross violations of human rights, acknowledging that some elements of criminal hate speech may classify as gross violations of human rights. After clarifying the general corporate remedial responsibility framework as covering modes of responsibility, remedial processes, and remedial outcomes, this Chapter clarifies that the remedial framework applies to online platforms that caused or contributed to criminal hate speech. This Chapter highlights the need for and proposes a corporate remedial responsibilities framework at the EU level, including for online platforms that caused or contributed to criminal hate speech. The proposed framework explores guarantees of non-repetition, restitution, and compensation as suitable remedial outcomes.

Chapter 6 presents the main findings related to the problem statement and research questions motivating this thesis, advances recommendations, and discusses areas of future research. This Chapter advances a comprehensive set of recommendations to strengthen the corporate human rights responsibilities of online platforms to counter criminal hate speech. These recommendations are addressed to three actors, i.e. legislators and policy makers, law enforcement bodies, and online platforms. Generally speaking, legislators and policy makers can rely more confidently on the general HRDD framework to conceptualize preventive, mitigating, and remedial responsibilities for online platforms to counter criminal hate speech. Law enforcement authorities should facilitate the establishment of reporting and investigative channels of criminal hate speech on online platforms. Online platforms should adhere to the HRDD framework and develop clear and transparent processes to prevent, mitigate, and remediate criminal hate speech disseminated through their services. The regulation of online platforms presents complex legal issues across different research disciplines, impacting a multitude of domestic and international jurisdictions and involving numerous stakeholders, including online platforms, public bodies, and individuals. Importantly, considering the constant changing nature of the services provided by online platforms, future research on measures to counter online hate speech require stronger human-rights centred interdisciplinary research.