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

Nave, E.V.R.

# 6 Conclusion[1]

This Chapter presents the main findings related to the problem statement motivating this thesis (Section 6.1), advances recommendations (Section 6.2), and discusses areas of future research (Section 6.3).

## 6.1 FINDINGS RELATED TO THE PROBLEM STATEMENT AND RESEARCH QUESTIONS

The main purpose of the thesis set out was (i) to conceptualize online hate speech, and (ii) to conduct a fundamental rights analysis of the ways in which online platforms perform their legal responsibilities in countering online hate speech.[2] Against this background, the problem statement motivating this thesis was formulated as follows:

> "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 responsibilities of online platforms to counter online hate speech whilst upholding fundamental rights?"

To address this problem statement, this study aimed to investigate legal pathways to strengthen the European regulatory and policy framework on the human rights responsibilities of online platforms to counter online hate speech. The short answer is two-fold. First, there is a need to strengthen, as a minimum legal standard, the human rights responsibilities of online platforms to counter the worst cases of hate speech, i.e. criminally actionable hate speech. Though this thesis focuses on the European standards distilling the

---

1 The findings in this Chapter were originally submitted as the NETHATE deliverable Number D2.2, titled "Tension between methods to counter 'hate speech' and the exercise of human rights", tasked to the NETHATE Early Stage Researcher Number 7. The objective prescribed in the NETHATE grant agreement for this deliverable was the development of a fundamental rights framework for analysing technological means to prevent hate speech. This objective corresponds to the research aim of this thesis. The NETHATE deliverables are published in CORDIS via DOI: 10.3030/861047. This Chapter was updated after the original submission. Cross-references should be read as referring to other references within the present Chapter.

2 NETHATE Grant Agreement, Annex 1, Description of the Action.

main elements of criminal hate speech, the standards deriving from international treaties are valid for an international conceptualization of criminal hate speech. The legal clarity on the main acts constituting criminal hate speech enables the establishment of more solid human rights responsibilities for online platforms.

Second, the framework regulating the human rights responsibilities of online platforms should expand on the responsibilities to prevent, mitigate, and remediate adverse impacts on human rights. As a minimum legal standard, the corporate human rights responsibility framework of online platforms should first and foremost focus on strengthening the responsibilities of platforms prone to higher systemic risks such as the proliferation of online hate speech. Currently, these platforms are very large online platforms, and particularly video-sharing platforms.[3] All standards proposed in this thesis must be explained to users in a way that they find effective, e.g. through clear notifications of changes in terms of services. Finally, the monitoring of the human rights framework developed in this thesis should be financed by the platforms themselves through the charging of a supervisory fee.[4]

The following subsections provide detailed explanations of the main findings to the four Research Questions by clarifying the legal framework conceptualizing hate speech and provides a working definition for the thesis (Section 6.1.1), and by exploring, respectively, the preventive (Section 6.1.2), mitigation (Section 6.1.3), and remedial (Section 6.1.4) corporate human rights responsibilities of online platforms to counter online hate speech. The aim is to develop a fundamental rights framework for analysing digital technologies used for countering online hate speech on online platforms by addressing the problem statement.

### 6.1.1 Legal conceptualization of hate speech

Currently, there is no legally binding definition of hate speech in international or European human rights law. The use of the term 'hate speech' with different connotations by several disciplines has contributed to its legal unclarity. From a legal perspective, hate speech refers to acts such as discrimination, threats, incitement to violence, to hatred, to genocide, to war crimes, to crimes against humanity, etc.[5] The basic legal framework regulating hate speech is found in human rights provisions including on the right to life, dignity, non-discrim-

3 Regarding the heightened responsibilities for video-sharing platforms, see Chapter 3, Section 3.4.2. on the EU Audiovisual Media Services Directive.

4 DSA, Art. 43.

5 Council of Europe, Committee of Ministers, Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech, Strasbourg, 20 May 2022.

ination, equality, participation in public life, expression, association. These provisions typically list categories that are expressly protected from discrimination, also known as 'protected characteristics' or 'impermissible grounds for hate speech'.[6] Depending on the human rights instrument, these characteristics can be formulated following an open-ended approach, and can include sex, gender, race,[7] colour, language, religion, political or other opinion, national or social origin, etc.

While hate speech is prohibited under international and regional human rights law, the effectiveness and operationalization of these prohibitions largely depend, at least in a first instance, on domestic implementation in national law. Importantly, there are significant discrepancies regarding the national transposition of such international human rights provisions. For example, though the United States of America (USA) ratified the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD),[8] in its Reservations to the Convention the USA did not accept to introduce any restriction on the right to freedom of speech, expression and association.[9]

These legal discrepancies are all the more relevant to address in the context of online hate speech on online platforms. Some of the biggest online platforms are based in the USA,[10] in China,[11] and in the United Arab Emirates,[12] and thus have typically followed the legal frameworks in these countries. As widely reported, many of these legal frameworks do not comply with international human rights and, as a result, online platforms have similarly applied content

6 Tarlach McGonagle (2011) 'Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas'. Following the work of McGonagle, this research employs "impermissible grounds" for hate speech as a way to refer to the traditionally called "protected characteristics" from discrimination. Some of the most common characteristics protected from discrimination based on human rights standards on non-discrimination include race, ethnicity, nationality, sex, gender, religion, disability. This research recognized that the expression "protected characteristics" can be understood as a legal condescending term that undermines the agency of people historically or systematically oppressed, and thus uses the expression "impermissible grounds" in an effort to depart from such patronizing approach. Applied to hate speech, "impermissible grounds" are the grounds based on which individuals are targeted by perpetrators of hate speech.

7 This research rejects theories of different human "races" as all humans belong to the same species. However, This research refers to "race" or "racialized" groups as a means to expose a colonial process whereby a dominant group ascribes to another group a racial identity for the purpose of continued oppression.

8 UN General Assembly, International Convention on the Elimination of All Forms of Racial Discrimination, United Nations, Treaty Series, vol. 660, p. 195, 21 December 1965.

9 United Nations Treaty Collection, International Convention on the Elimination of All Forms of Discrimination, Declarations and Reservations available at <https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-2&chapter=4&clang=_en#EndDec> accessed 21 Feb 2024.

10 *E.g.* Facebook, YouTube, WhatsApp, Instagram.

11 *E.g.* TikTok, WeChat.

12 *E.g.* Telegram.

regulation standards that are not human rights compliant.[13] A specific example of content regulation by online platforms that did not comply with human rights standards was Facebook's definition of categories to be protected from hate speech. Relying on a standard that favoured the protection of the majority, Facebook removed a post suggesting that "all white people were racist" but authorized a post incentivizing the "killing of radicalized Muslims."[14] To the extent that these online platforms cater to a European user base, they must comply with European human rights standards. In this context, this Chapter's Research Question was:

> *What are the main elements of hate speech under European human rights standards, do they align with the conceptualization of hate speech by critical legal theory, and to what extent do they require further clarification?*

The methodology employed in this Chapter was doctrinal and interdisciplinary legal research. Doctrinal research focusing on applicable legal frameworks to online hate speech in Europe sought to clarify the existing legal standards and to identify and address legal loopholes. Interdisciplinary legal research aims to investigate the interplay between European human rights law and critical legal (race) theory and (black) feminist intersectionality theory. These last two theoretical frameworks were selected as these gave prominence to the term (racist) "hate speech" in legal scholarship.

To answer this Research Question, the legal foundations of the conceptualization of hate speech by critical race theory were assessed. Building on the work of Matsuda and Gelber, followed the approach that all types of hate speech are used to perpetuate a relationship of power imbalance and to target historically or systematically oppressed groups.[15] Similarly, building on the work of Crenshaw, this Chapter emphasized that when a person's lived experiences lie at the intersection of various systems of oppression (race, gender, queerness, ableism, etc), this intersectionality plays as an factor that

13 E.g. Human Rights Watch (2020) Big Tech's Heavy Hand Around the Globe, Facebook and Google's dominance of developing-world markets has had catastrophic effects. US regulators should take note available at <https://www.hrw.org/news/2020/09/08/big-techs-heavy-hand-around-globe> accessed 28 August 2024.

14 Julia Angwin, ProPublica & Hannes Grassegger, Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children, PROPUBLICA (June 28, 2017), available at <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms> accessed 21 Feb 2024.

15 Mari J. Matsuda et al, Words that wound: Critical race theory, assaultive speech, and the first amendment 6 (2018) Id., 16; Katharine Gelber (2021) "Differentiating hate speech: a systemic discrimination approach." Critical Review of International Social and Political Philosophy, 24(4), 393–414 available at <https://doi.org/10.1080/13698230.2019.1576006> accessed 28 August 2024.

aggravates the harm caused by hate speech.[16] Finally, the analysis presented in this Chapter also underlined the cumulative effects of being targeted by hate speech as a factor enhancing the harm in hate speech.[17] These are the main elements conceptualizing hate speech stemming from critical legal theory.

This Chapter then evaluated the main elements of hate speech under European human rights standards, both at the Council of Europe and the European Union, and their alignment with the conceptualization of hate speech by critical scholars. This analysis focused on legal and policy instruments that conceptualize hate speech from a substantive regulatory perspective.[18]

At the Council of Europe level, this Chapter assessed the European Convention on Human Rights (ECHR), as interpreted by the European Court of Human Rights (ECtHR) in its case-law, and other treaties regulating hate speech, as well as non-treaty initiatives. The analysis in this Chapter found that the ECtHR acknowledges, in its jurisprudence that, in hate speech cases, it is important to consider the political and social context, as well as the victims' perspective. Nevertheless, though these elements align with the historical and intersectional elements of oppression in hate speech alluded to by critical scholars, the ECtHR fails to formally and consistently address them. Overall, it is positive to note that there is a growing body of legal and policy instruments at the Council of Europe that adopt open-ended conceptualizations of the impermissible grounds for hate speech. This development shows increased alignment with the theory of the intersectionality of systems of oppression advanced by critical legal scholars. Examples of such instruments include the Istanbul Convention,[19] and the Recommendation CM/Rec(2022)16.[20] Notably, Recommendation CM/Rec(2022)16 is the most comprehensive standard-setting instrument clarifying human rights standards for the regulation of hate speech both offline and online and is an instrumental reference in this thesis. Nonetheless, the legal and policy framework at the Council of Europe fail to consistently mention that a key element in the conceptualization of hate speech is its utilization to perpetuate historical or systematic systems of oppression.

16 Devon W. Carbado , Kimberlé Williams Crenshaw , Vickie M. Mays and Barbara Tomlinson (2013). Intersectionality, Du Bois Review, 10(2), 303–312 available at <https://doi.org/10.1017/S1742058X13000349> accessed 28 August 2024.

17 Richard Delgado & Jean Stefancic (2004) "Understanding Words That Wound", 29 (4) 917–918.

18 Instruments covering procedural regulation of the responsibilities and liabilities of stakeholders involved in the regulation of hate speech is dealt with in Chapters III to V (inclusive) of this thesis.

19 Convention on Preventing and Combating Violence Against Women and Domestic Violence, May 11, 2011, E.T.S. 210, available at <https://rm. coe.int/168008482e> accessed 21 Feb 2024.

20 Council of Europe, Committee of Ministers, Recommendation CM/Rec(2022)16.

At the European Union level, this Chapter investigated general principles, primary sources[21] such as the Charter for Fundamental Rights of the European Union (CFREU), and secondary sources such as the Council Framework Decision on combating certain forms and expression of racism and xenophobia by means of criminal law (Framework Decision),[22] the Audiovisual Media Services Directive (AVMSD). Overall, the analysis in this Chapter found that there is a lack of consistency across relevant instruments at the European Union level conceptualizing hate speech in the way that they approach the impermissible grounds for hate speech. For example, while the AVMSD adopts an open-ended approach, the Framework Decision limits its scope to 'race, colour, religion, descent or national ethnic origin'. This results in the lack of a consistent legal framework to protect communities often targeted by online hate speech, such as the queer community. Hence, the initiatives at the European Union level fail to expressly and consistently acknowledge the historical, systematic, and intersectional elements of hate speech conceptualized by critical race scholars. The European Commission has initiated a legislative process calling for a Council Decision extending the list of 'EU-crimes 'as per Article 83 Treaty on the Functioning of the European Union to hate crime and hate speech.[23] This initiative has the goal of addressing the loophole regarding the different approaches for impermissible grounds, and of clarifying the state obligations in ensuring protection from hate speech. Notwithstanding, while this European Union legislative initiative is not adopted and thus does not establish binding obligations for Member States, Recommendation CM/Rec(2022)16 provides the most comprehensive and up-to-date standard-setting framework for the regulation of hate speech in the European context.

Finally, this Chapter defended that the conceptualization of hate speech in the European context could benefit from further clarification to align with the conceptualization of hate speech by critical legal theory and proposes five guiding principles contributing for a better alignment. First, people targeted by hate speech have been or are either historically or systematically oppressed. This Chapter emphasized how the neutral conceptualization of hate speech

21 European Parliament, Fact Sheets on the European Union, Sources and scope of European Union Law, available at <https://www.europarl.europa.eu/factsheets/en/sheet/6/sources-and-scopeof-european-union-law> accessed 28 August 2024.

22 European Union, Council Framework Decision 2008/913/JHA.

23 European Commission, Extending EU crimes to hate speech and hate crime, COM(2021) 777 final, available at <https://eur-lex.europa.eu/resource.html?uri=cellar:4d768741-58d3-11ec-91ac-01aa75ed71a1.0002.02/DOC_1&format=PDF> and available at <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/extending-eu-crimes-hate-speech-and-hate-crime_en> accessed 28 August 2024.

by European frameworks hinders the effective regulation of hate speech.[24] Second, hate speech is always illegal, either under civil or administrative law or under criminal law. Third, hate speech should only be criminalized in its most serious forms. Fourth, the following 'contextual variables' should be used to assess the severity of hate speech: political and social background; intent of the speaker; speaker's status or role in society; content of the expression; extent and reach of the expression; and the nature of the audience.[25] In this exercise, it is important to explicitly account for socio-historical records of oppression and the victims' potential intersectional position between various oppressive systems. Fifth, in the context of online hate speech where content typically spreads through large audiences, the contextual variable of reach must be carefully considered and attributed an inherently increased weigh than hate speech in offline settings.

### 6.1.2 Human rights responsibilities of online platforms to prevent criminal hate speech

Chapter 3 is the first of three studies that, building on the conceptualization of criminal hate speech advanced in Chapter 2, focus on the human rights responsibilities of online platforms to counter online hate speech. This Chapter delved deeper into the preventive human rights responsibilities of online platforms to conceptualize hate speech based on human rights standards. More specifically, this Chapter's Research Question was:

> *To what extent is there a legal standard emanating from the European human rights preventive due diligence framework prescribing the responsibility for online platforms to align their terms of service, as a minimum legal standard, with the conceptualisation of the criminal hate speech as explained in the European human rights standards, in particular with the Recommendation CM/Rec(2022)16?*

To clarify the conceptualization of criminal hate speech stemming from European standards, this Chapter employed doctrinal research. The most relevant instrument in this context is Recommendation CM/Rec(2022)16 which in Paragraph 11 summarizes the key hate speech acts that are criminally action-

24 Within the current European legal and policy frameworks, a white, heteronormative, cisgender, neurotypical, and abled men, thus belonging to various privileged societal groups, could in principle claim to be a victim of hate speech. This would render the legal framework regulating hate speech ineffective in delivering on its goal to halt hate speech towards historically or systematically oppressed groups.

25 Michel Rosenfeld (2002) Hate Speech in Constitutional Jurisprudence: A Comparative Analysis Conference: The Inaugural Conference of the Floersheimer Center for Constitutional Democracy: Fundamentalisms, Equalities, and the Challenge to Tolerance in a Post-9/11 Environment, 24 CARDOZO L. REV. 1523, 1565.

able based on existing treaty obligations. These broadly include: incitement to genocide, violence, or discrimination; threats; public denial, trivialization and condoning of genocide, crimes against humanity or war crimes; and, intentional dissemination of material with these expressions.[26] This Chapter adopted a critical approach of this Paragraph by highlighting the need to consider the intersectionality of historical or systematic systems of oppression in criminal hate speech.

To clarify the European human rights preventive due diligence framework applicable to online platforms, this Chapter continued to employ doctrinal research to analyse Principle 15 of the United Nations Guiding Principles on Business and Human Rights (UNGPs),[27] and the legally binding Corporate Sustainability Due Diligence Directive (CSDDD)[28] adopted by the European Union. This analysis clarified that any business, including online platforms, should comply with the broader framework requiring business to respect to human rights. Hence, online platforms must adopt a policy commitment to respect human rights. Delving deeper into sector specific human rights responsibilities for online platforms stemming from European standards, this Chapter suggested that the terms of service should be considered the adequate place for conveying the policy comment towards human rights. By examining the Digital Services Act (DSA)[29] and the AVMSD,[30] this Chapter claimed that online platforms should align their terms of service with the conceptualisation of criminal hate speech and of corporate human rights responsibilities as explained in the European human rights standards.

After that, this Chapter employed comparative research to present three case studies of online platforms' lack of compliance with human rights standards in the conceptualization of hate speech.[31] These case studies show that Facebook (Meta Platforms, Inc.), X Corp. (previously Twitter, Inc.), and YouTube, despite prohibiting hate speech on their terms of service, all fail to clearly identify hate speech that derives from human rights treaty obligations and that is criminally actionable. Moreover, none of these platforms recognizes their responsibilities to align with human rights policies, and due diligence and remedial processes.

This Chapter suggested an innovative human rights standard. Building on emerging corporate human rights instruments clarifying the responsibilities of online platforms in the European Union, this Chapter advanced a new legal standard that online platforms, and particularly very large online platforms,

26 CM/Rec(2022)16, Paragraph 11.
27 UN Human Rights Council (2011) 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie A/HRC/17/31.
28 European Union, CSDDD.
29 European Union, DSA, Article 93.
30 European Union, AVMSD.
31 NETHATE Grant Agreement, Annex 1, Description of the Action.

video-sharing platforms, and platforms under the scope of the CSDDD, should align their terms of service with the conceptualization of criminal hate speech deriving from international and regional human rights standards. Additionally, this Chapter also suggested that terms of service should explicitly align with the human rights due diligence process to prohibit, remove and report criminal hate speech to law enforcement.

### 6.1.3 Human rights responsibilities of online platforms to mitigate criminal hate speech on end-to-end encrypted services

Chapter 4 explores the human rights responsibilities of online platforms to mitigate hate speech in the specific case of end-to-end encrypted (E2EE) services. The proliferation of hate speech on online platforms was initially reported on open-ended encryption environments such as news feeds of public accounts and public comments on platforms accessible upon the creation of a user account. Over time, online platforms started to increase and improve privacy settings. On the one hand, these privacy settings empower for example human rights activists to organize and express themselves protecting them from prosecution by autocratic states.[32] On the other hand, studies alert to the fact that the increased anonymity can lead to higher criminal activity associated with decreased accountability.[33] One such type of environment is E2EE messaging services, which is the focus of this Chapter's research.

In recent years, the migration of hate mongers to E2EE services has posed new regulatory challenges.[34] In particular, legislators and law enforcement have battled with advancing the human rights responsibilities of online platforms providing E2EE services to counter hate speech without compromising freedom of expression, association, privacy and data protection rights.[35] In this context, this Chapter's Research Question was:

> *To what extent can an innovative legal interpretation of technological developments clarify and expand the human rights due diligence (HRDD) of online platforms providing end-to-end encrypted (E2EE) services in the European context*

32 Amnesty International (20216) Encryption A Matter of Human Rights, available at <https://www.amnesty.nl/content/uploads/2016/03/160322_encryption_-_a_matter_of_human_rights_-_def.pdf> accessed 7 September 2023.

33 EUROJUST (2021), Third report of the observatory function on encryption, available at <https://www.eurojust.europa.eu/publication/third-report-observatory-function-encryption> accessed 21 February 2024.

34 Center for Democracy &Technology (2021) Outside looking In – Approaches to Content Moderation in End-to-End Encrypted Systems, available at <https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf> accessed 21 Feb 2024.

35 Center for Democracy &Technology (n 34).

*to not host criminal hate speech in the form of incitement to violence? If so, can this innovative interpretation result in new HRDD responsibility standards for E2EE services' cooperation with law enforcement?*

To answer this Research Question, this Chapter employed an interdisciplinary human rights doctrinal analysis of innovative technologies used for regulation of hate speech in E2EE services provided by online platforms. The scope of this research is limited to elements of hate speech inciting to violence towards historically or systematically oppressed people. This scope is justified based on the potential harm caused by online hate speech on E2EE services. To clarify, on open-ended encryption digital environments, networks of users can be composed of hate speech perpetrators and victims and thus online hate speech can be directly targeted at its victims. Contrarily to this, on E2EE services the networks are typically composed of likeminded people and thus hate speech perpetrators and victims would not engage directly. Hence, the harm of hate speech on E2EE services is to radicalize hate speech perpetrators by inciting them to violence. Additionally, the number of participants in such E2EE messaging services is increasing, reaching the thousands of users in just one chat. This poses a higher concern as, presently, large groups can more easily than ever incite to hatred and violence, potentially leading to offline crimes while facing little or no accountability. Against this background, this Chapter analysed, first, the human rights responsibilities of online platforms providing E2EE services and, second, the technological advancements in content regulation on E2EE services.

First, building on the broader business and human rights framework deriving from the UNGPs and the CSDDD, this Chapter explained that all businesses must mitigate adverse impacts that are directly linked to their operations, products or services, including those businesses providing E2EE services. Additionally, under the CSDDD, this Chapter clarified that businesses with 500+ employees and a turnover of over € 150 million worldwide, such as Facebook and WhatsApp, have to *inter alia* protect the right to life and security, prohibition of cruel, inhuman or degrading treatment.[36] This Chapter posited that this responsibility applies to the E2EE services provided by online platforms, and especially to very large online platforms.

This Chapter found that, within the EU sector-specific framework regulating online platforms, the E2EE services arguably fall within the scope of the DSA in two ways. If regarded as a type of service provided by online platforms within the DSA criteria, E2EE messaging services provided by such online platforms would have to comply with the DSA human rights requirements

36 European Union, CSDDD, Annex I.

(*e.g.* Facebook Messenger and X).[37] Alternatively, this Chapter explained that platforms that by nature only provide E2EE services and that include the option of open or public channels,[38] would likewise fall within the remit of the DSA directly (*e.g.* WhatsApp).[39] Further to this, this Chapter defended that the human rights responsibilities prescribed by the AVMSD arguably apply to E2EE services to the extent that they hold editorial responsibility for the display order of Graphics Interchange Format (also known as GIFs). Additionally, this Chapter highlighted that online platforms signatories to the EU Code of Conduct to counter illegal hate speech online, some providing E2EE services, must comply with human rights. These include for example, Facebook Messenger, Snapchat, Viber, and X's encrypted messaging services. Finally, this Chapter reiterated the key standard-setting Recommendation CM/Rec(2022)16, which underlines that internet intermediaries should comply with human rights due diligence processes independently from size, sector, operational context, nature, etc.

Second, this Chapter defended that the application of the human rights due diligence responsibilities to E2EE services depends on the available technological advancements. In this context, this Chapter advanced a regulated application of metadata, hashing and homomorphic encryption enabling the deployment of disruption techniques to mitigate incitement to violence in open groups or large groups on E2EE services. This Chapter found that this regulatory standard provides an adequate balance between the protection from being harmed as a result of incitement to violence and the protection of the rights to freedom of expression, assembly, privacy, and data protection. This Chapter advanced the debate on the regulation of metadata in a way that is compliant with the GDPR and with the e-Privacy Directive.[40]

To summarize, this Chapter proposed an innovative corporate human rights responsibility to mitigate incitement to violence on E2EE services. This Chapter claimed that criminal hate speech in the form of incitement to violence shared in E2EE services in open or large groups meets the highest threshold of risks to human rights. Additionally, online platforms providing E2EE services, and in particular very large online platforms, can be associated with increased systemic risks to human rights as privacy-preserving features may increase criminal activity. As such, online platforms, and especially very large online

37 Facebook Help Center, What end-to-end encryption on Messenger means and how it works, available at <https://www.facebook.com/help/messenger-app/786613221989782?cms_id=786613221989782> accessed 21 February 2024.

38 European Union, DSA, Recital 14.

39 E.g. The current features of WhatsApp groups arguably characterized these settings as open channels given the accessibility for members to join. *E.g.* available at <https://www.whtsgrouplinks.com/> accessed 6 February 2024.

40 Currently, metadata is not regulated at the EU level which results in a worrying legal vacuum where compliance with corporate human rights responsibilities are not monitored.

platforms, providing E2EE services have heightened responsibilities[41] to mitigate incitement to violence facilitated by their services. Restrictions on the right to freedom of expression must comply with the legal requirements in Article 10(2) ECHR and be the least intrusive means possible. This Chapter developed a standard that would arguably be one of the least intrusive methods to counter criminal hate speech on E2EE services. Briefly, following the creation of a database, translated into all languages active in a given platform, of incitement to violence targeting historically or systematically oppressed groups, metadata could be used to monitor the size of the group. Above a given threshold of group size to be decided based on the state-of-art, hashing and homomorphic encryption could be employed to detect known text or image content matching the database. Once text or image is detected, disruption techniques can be employed such as freezing or division of the large group. This Chapter advocated that such a standard would protect the user's identity and enable the platform providing E2EE services to disrupt groups inciting to violence in a human rights compliant manner.

### 6.1.4 Human rights responsibilities of online platforms to remediate criminal hate speech

Chapter 5 explores the remedial human rights responsibilities of online platforms which caused or contributed to online hate speech. There is an increasing number of reports by human rights organizations alerting to the implications of online platforms, and particularly of very large online platforms, in hosting and spreading online hate speech. For example, the United Nations and Amnesty International revealed that Meta played a significant role contributing to the genocide of the Rohingya in Myanmar after its algorithm not only failed to remove but also amplified hate speech towards this community.[42] In this case, Amnesty's investigation uncovers that the company knew that its algorithm contributed to the rise of extremism on the platform. Moreover, an internal company document titled "Facebook and Responsibility" shows that Facebook itself recognized that for instance its ranking algorithm makes it responsible for any harm caused by exposure to said ranked content.[43] Nevertheless, following a demand for an effective remedy by the Rohingya requesting Facebook to fund a USD 1 million education project, Meta refused, saying that the proposal was not directly linked to its product.[44] This example illus-

41 European Union, DSA, and Council of Europe, Committee of Ministers, Recommendation CM/Rec(2022)16.

42 Amnesty International, 'Myanmar: The social atrocity : Meta and the right to remedy for the Rohingya' (2022) https://www.amnesty.org/en/documents/ASA16/5933/2022/en/ (accessed 28 May 2024), 6.

43 Amnesty International (n 42), 43.

44 Amnesty International (n 42), 66.

trates how very large online platforms are not complying with their human rights remedial responsibilities.

In this context, there was a need to research the specific application of the framework of corporate remedial responsibilities for online platforms. The general international business and human rights framework articulates the businesses responsibility to provide for remediation mechanisms of any adverse impact on human rights that the business caused or contributed to. However, the legal framework recently developed at the EU level regulating online platforms fails to clarify the remedial responsibilities of these businesses. Against this background, this Chapter's Research Question was:

> *To ensure the right to an effective remedy, how can European Union and Council of Europe legislators better align the legal framework on the corporate remedial responsibilities of online platforms which caused or contributed to criminal hate speech with the general framework on corporate remedial responsibilities?*

To answer the Research Question, this Chapter employed doctrinal research to review the criminally actionable hate speech as per Recommendation CM/Rec(2022)16 and defend the possibility that elements of criminal hate speech may amount to gross violations of human rights and thus result in the application of the right to remedy and reparation for victims of gross human rights violations.

This Chapter then defended that online hate speech on online platforms can cause psychological, physical, and economic harms,[45] as well as that the continued exposure to hate speech (also referred as the cumulative effect) reflects an aggravating factor heightening the harm caused by hate speech.[46] In the context of criminal hate speech cases amounting to gross violations of human rights and disseminated through online platforms, and especially through very large online platforms, this Chapter advocated that the European legal framework, both at the European Union and at the Council of Europe, should recognize the increased degree of harm when compared to other cases of hate speech. In light of Article 13 of the ECHR, which establishes the right to an effective remedy before a national authority, this Chapter underlined the States' duty to investigate and ensure "diligent, thorough, and effective" access to an effective remedy for people targeted by criminal hate speech, including by that amounting gross violations of human rights. This Chapter

45 This position builds on the conceptualization of harm advanced by critical race and black feminist scholars as explained in Eva Nave "Hate Speech, Historical Oppressions, and European Human Rights." Buff. Hum. Rts. L. Rev. 29 (2022): 83, p. 91.

46 Richard Delgado (1982) 'Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling', 17 Harvard Civil Rights Liberties Law Review 133.

further explained the States' duty to ensure the right to an effective remedy including when violations are attributable to businesses.[47]

This Chapter explained that the general framework on corporate human rights responsibilities, deriving from the UNGPs and from the now legally binding CSDDD,[48] assigns remedial responsibilities for businesses that caused or contributed to actual adverse impacts on human rights or to potential adverse impacts on human rights that are unavoidable.[49] This Chapter further clarified that, for cases where the business caused or contributed to gross human rights violations, businesses have the responsibility to adopt remediation processes that legitimately, promptly, and effectively repair the gross human rights violation.[50] This Chapter emphasized that, while the effectiveness of a remediation process must be evaluated by the victims themselves, the general framework for effective remedies includes: restitution; satisfaction; compensation; rehabilitation; and, guarantees of non-repetition.[51]

In evaluating the corporate remedial responsibilities of online platforms, this Chapter focused on the European Union legal framework given the recent adoption of binding legislation on both platform and artificial intelligence governance. This Chapter explained that the DSA underlines the importance that online platforms, with an emphasis on very large online platforms, comply with human rights, including with the right to an effective remedy. Furthermore, this Chapter interpreted Article 5 of the AI Act as a prohibition for online platforms to deploy content regulation techniques that amplify criminal hate speech,[52] and advocates that a breach of this corporate human rights responsibility by online platforms should result in remedial responsibilities.

To summarize, this Chapter advanced a legal approach to strengthen the EU framework of human rights responsibilities of online platforms. More specifically, this Chapter reconciled the individual right to an effective remedy, the States' duty to ensure the respect for the right to remedy, and the remedial

47 Council of Europe, Freedom of Expression, Effective Remedies, Explanatory Memorandum, available at <https://www.coe.int/en/web/freedom-expression/effective-remedies-explanatory-memo> accessed 28 August 2024.

48 European Union, CSDDD, Art. 3(1)(l).

49 United Nations, UNGPs (n 27), Guiding Principle 15.

50 United Nations, Human Rights Office of the High Commissioner, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, A/RES/60/147, available at <https://www.ohchr.org/en/instruments-mechanisms/instruments/basic-principles-and-guidelines-right-remedy-and-reparation> accessed 28 August 2024, Article 11(b).

51 United Nations, Human Rights Office of the High Commissioner, The Corporate Responsibility to Respect Human Rights, An Interpretative Guide, available at <https://www.ohchr.org/sites/default/files/Documents/publications/hr.puB.12.2_en.pdf>, Q. 64.

52 European Union, AI Act. Article 5 establishes the prohibition of AI systems that "deploy subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm".

responsibilities of online platforms that caused or contributed to criminal hate speech. Building on the general international corporate remedial responsibilities standards, this Chapter clarified that the corporate remedial responsibilities framework at the EU level of online platforms should have to cover: modes of responsibilities; remedial processes; and remedial outcomes. The modes of responsibility should have to clarify that an online platform that caused or contributed to criminal hate speech has direct corporate remedial responsibilities. This framework is all the more relevant in the case of very large online platforms.

This Chapter concluded by defending that corporate remedial outcomes by online platforms should have to be tailored to the harm caused by criminal hate speech, particularly that amounting to gross human rights violations, and presents tailored remedial outcomes: restitution and satisfaction; compensation and rehabilitation; and guarantees of non-repetition. While recognizing the various types of harms caused by criminal hate speech on online platforms, this Chapter focused on the specific harm of constrained participation and, in this context, advances a legal remedy through which platforms amplify the survivors' speech. Finally, this Chapter advocated for platforms to comply with the guarantee of non-repetition they should have to ensure that their business model prioritizes human rights over profit.

## 6.2 RECOMMENDATIONS

This section advances recommendations for three actors, i.e. legislators and policy makers, law enforcement bodies, and for online platforms. These recommendations are presented in four themes, each corresponding to the one of the main substantive Chapters, i.e. Chapters II to V.

Regarding the standards for the legal conceptualization of hate speech in the European context, Chapter 2 recommends that the European regulatory and policy framework should explicitly acknowledge the conceptualization of hate speech by critical legal scholars as expressions intended to perpetuate historical or systematic oppression. Moreover, it should expressly account for the intersectionality of systems of oppression as well as the cumulative effect of continued exposure to hate speech as aggravating factors harming people targeted by hate speech. In addition, in the context of online hate speech, the contextual variable of reach must by default be carefully considered as a potential aggravating factor for the harm caused by hate speech.

Specific avenues for the integration of these three standards can be recommended. At the level of the Council of Europe, a starting point could be through the jurisprudence of the ECtHR. To be more specific, following the adoption of CM/Rec(2022)16 by the Committee of Ministers on Combating Hate Speech, the ECtHR has the opportunity to apply and clarify the application of the standards in CM/Rec(2022)16. One way to do this, could be by

clarifying that CM/Rec(2022)16 follows the three standards recommended here for the legal conceptualization of hate speech in Europe to better align with critical legal standards. At the European Union level, should the Council of the European Union follow the European Commission Communication to adopt a Decision extending Art. 83 of the TFEU to hate crime and hate speech, this would be the adequate instrument to reflect the three standards advanced in this research.

Regarding the human rights responsibilities of online platforms to *prevent* criminal hate speech, Chapter 3 presents recommendations for all three actors. First, addressing legislators and policy makers, Chapter 3 recommends that the European Commission should issue a best practice under Article 35(3) of the DSA and Article 13 of the CSDDD indicating that: i) very large online platforms, and particularly video-sharing platforms, should explicitly mention in their terms of service that they prohibit, remove, archive, and report to law enforcement criminal hate speech in line with Paragraph 11 of the Recommendation CM/Rec(2022)16; and, ii) very large online platforms, and particularly video-sharing platforms, should explicitly mention in their terms of service how their content regulation processes (including but not limited to content moderation, ranking, and recommendation algorithms) align with the human rights due diligence processes.

Second, addressing law enforcement bodies, Chapter 3 recommends, as a minimum legal standard, the establishment of mechanisms for investigating online criminal hate speech in line with Paragraph 11 of the Recommendation CM/Rec(2022)16 and acknowledging the intersectionality of historical or systematic systems of oppression perpetuated by hate speech.

Third, addressing online platforms, Chapter 3 recommends that these should explicitly mention in their terms of service that they prohibit, remove, archive, and report to law enforcement criminal hate speech in line with Paragraph 11 of the Recommendation CM/Rec(2022)16 as well as how their content regulation processes align with human right due diligence processes. As this best practice would contribute to reporting criminal offences to law enforcement and would thus classify as a high-risk AI system under Article 6(a) of the AI Act, online platforms would be required to comply with enhanced human rights due diligence standards throughout the application of this standard.

Regarding the human rights responsibilities of online platforms to *mitigate* criminal hate speech on E2EE services, Chapter 4 also presents recommendations for all three actors. First, addressing legislators and policy makers, Chapter 4 recommends a new standard expanding the human rights responsibility of online platforms providing E2EE services to mitigate by disruption incitement to violence on open or large channels. The key legal objective would be to protect the right to life and safety. This legal standard suggests an innovative interpretation of metadata, hashing, and homomorphic encryption. To this end, Chapter 4 recommends that legislators and policy makers work

with civil society and human rights activists for the creation and translation of a database of "incitement to violence" which should adopt a very strict interpretation aligned with the acknowledgement of the intersectionality of historical or systematic systems of oppression perpetuated by hate speech.

It is recommended that either national or administrative authorities issue an order requiring online platforms providing E2EE services to comply with this legal standard. Such a legal order could have legal grounding in Article 9 of the DSA and Article 6 of the GDPR expanding the human rights responsibility of online platforms providing E2EE services to mitigate by disruption incitement to violence on open or large channels utilizing metadata, hashing, and homomorphic encryption. Similar to the recommendation in Chapter 3, also here the suggested standard would qualify as a high-risk AI system and its deployment would have to undergo strict human rights due diligence processes as per the AI Act. Finally, through this standard, this Chapter advocates for the first regulation of metadata, in compliance with the GDPR and with the e-Privacy Directive.[53]

Second, addressing law enforcement bodies, Chapter 4 recommends, as a minimum legal standard, the establishment of mechanisms to follow up on online platforms reporting online incitement to violence to investigate offline incitement to violence targeting historical or systematic oppressed communities. Third, addressing online platforms, Chapter 4 recommends that these should fund the creation and translation of the public database of "incitement to violence" as well as comply with human rights responsibility to mitigate incitement to violence by deploying metadata, hashing, and homomorphic encryption and the respective human rights safeguards. As this best practice would contribute to reporting criminal offences to law enforcement and would thus classify as a high-risk AI system under Article 6(a) of the AI Act, online platforms would be required to comply with enhanced human rights due diligence standards.

Regarding the human rights responsibilities of online platforms to *remediate* the harm that they caused or significantly contributed to by amplifying criminal hate speech, Chapter 5 presents recommendations to legal and policy makers as well as to online platforms. First, this Chapter recommends that the European Commission issues a detailed guidance on the operationalization of Article 21 of the DSA in alignment with the general corporate remedial responsibilities framework stemming from the UNGPs and from the legally binding CSDDD. Essential aspects that should feature in this guidance cover: i) modes of responsibilities, with increased remedial responsibilities for cases of criminal hate speech amounting to gross human rights violations; ii) minimum standards for remedial processes covering required compliance with legitimacy, promptness, and impartiality criteria; and, iii) minimum standards

53 Currently, metadata is not regulated at the EU level which results in a worrying legal vacuum where compliance with corporate human rights responsibilities are not monitored.

for remedial outcomes tailored to effectively address the harm caused by criminal hate speech.

Chapter 5 recommends that such detailed guidance could take the form of a new the European Union legal or policy instrument providing a comprehensive overview of remedial responsibilities of online platforms that caused or contributed to content deemed illegal in the European Union, particularly criminal hate speech. Such an instrument should be accompanied by a monitoring mechanism and dedicated monitoring team operating with the European Commission Directorate General Connect, more specifically within the Digital Services Act Enforcement Team.[54]

Second, addressing online platforms, Chapter 5 recommends that online platforms which caused or significantly contributed to criminal hate speech should establish legitimate, effective, and impartial remedial mechanisms. This Chapter recommends heightened human rights responsibilities for online platforms which caused or contributed to criminal hate speech amounting to gross human rights violations.

Third, Chapter 5 recommends that online platforms should have to comply with minimum human rights standards covering remedial outcomes tailored to effectively address the harm caused by criminal hate speech. This Chapter advances that a specific remedial outcome to provide for restitution of the harm caused by online platforms which amplified criminal hate speech could encompass a deliberate amplification of content portraying the narrative of the people targeted by hate speech to introduce affirmation speech policies in their content ranking, moderation, and recommendation algorithms. To conclude, this Chapter defends that for online platforms to comply with the remedial outcome of guarantees of non-repetition, it is critical to comply with Article 5 of the AI Act and review all facets of content regulation, including moderation, ranking and recommendation algorithms to ensure that the business model prioritizes human rights over engagement and profit.

## 6.3 Areas for future research

This section reviews the main areas for future research, first, regarding the overall thesis and, second, regarding the specific Chapters II to V. In respect of the main areas of future research related to the thesis, three research areas stand out. One main area for future research relates to the importance of studying the balance between the power of online platforms, public bodies, and individuals in regulatory initiatives. As more regulation is developed by public bodies to ensure the compliance of online platforms with human rights,

54 European Commission, Communications Networks, Content and Technology, available at <https://commission.europa.eu/about-european-commission/departments-and-executive-agencies/communications-networks-content-and-technology_en> accessed 26 August 2024.

it is crucial that this does not lead to a simple transfer of power from platforms to public administration without involving and empowering individuals and civil society along the regulatory process.[55]

The second main area requiring further research relates to the study of the regulation of online platforms' responsibility and human rights safeguards required to collaborate with law enforcement bodies. In assessing the role of law enforcement, it is essential to acknowledge that these entities are fallible and can suffer from infiltration by violent extremists.[56] This area of future research should also explore the application of monitoring systems to the utilization of online platforms by law enforcement bodies and should, overall, cater for a human rights compliant collaboration between online platforms and law enforcement.

The third main area for future research originating from this thesis is the need to better investigate the interplay between different fields of law to ensure the best possible design of legislation to regulate online platforms. Future research on this topic should comprise the combined study of businesses and human rights, platform and AI governance, computer science, as well as regulation applicable to the specific type of illegal content. In isolation, these fields do not clarify the applicable regulation to online platforms, however, a combined approach of these research areas creates a pathway for more solid and practice informed regulation.

Analysing the specific areas of future research stemming from Chapter 2, this thesis suggests further research regarding the positionality and lived experiences of the researchers in relation to the topic, particularly those researchers in the field of hate speech studies. It is important to invest resources to challenge the typical exclusionary setting that is academia in Europe where, as mentioned by El-Tayeb, various systems of marginalization, and consequently privilege, are endorsed and perpetuated.[57]

Chapter 3 emphasizes the need for further specific research to examine the human rights responsibilities of online platforms beyond the preventive measures to counter criminal hate speech. For example, further research is needed to understand the most adequate public oversight mechanisms monitor-

55 Martin Husovec (2024) "Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation of Global Internet Rules." *Berkeley Technology Law Journal* 38.3.

56 Aurelien Mondon and Aaron Winter (2020) "Reactionary democracy: How racism and the populist far right became mainstream." Verso Books.

57 See, e.g., Fatima El-Tayeb (2011) European Others, Queering Ethnicity in Postnational Europe, 229; William E. Donald. (2024) Merit beyond metrics: Redefining the value of higher education. Industry and Higher Education; Robin Cowan, Moritz Müller, Alan Kirman, Helena Barnard, Overcoming a legacy of racial discrimination: competing policy goals in South African academia, Socio-Economic Review, Volume 22, Issue 3, July 2024, Pages 1413–1449, https://doi.org/10.1093/ser/mwad043; Williams, M. T. (2019) Adverse racial climates in academia: Conceptualization, interventions, and call to action. *New ideas in psychology*. 5558–67; Llorens, A. et al. (2021) Gender bias in academia: A lifetime problem that needs solutions. *Neuron (Cambridge, Mass.)*. 109 (13), 2047–2074.

ing the compliance with the removal, archiving, and reporting to law enforcements of criminal hate speech. In a context where online platforms are used by victims, survivors, bystanders, and perpetrators of hate speech, including criminal hate speech and that amounting to gross human rights violations, it is imperative to regulate the responsibility of online platforms to archive content which can later be used as evidence in criminal prosecutions.

There are three specific areas for future research identified in Chapter 4. The first relates to the need to further analyse the role of linguistics to counter hate speech on online platforms, in particular linguistics informed by the lived experiences of the community targeted by hate speech. As platforms adopt datasets of interpretations of text, images, and audio from posts and subsequently train their content regulation algorithms to identify such content, it is essential to develop translations for such datasets of interpretations into all known active languages on their services.[58]

A second specific area for further research identified in Chapter 4 relates to the implementation of the proposed regulatory standard on the operationalization of disruption strategies on E2EE services. This thesis acknowledges that such a standard may detect speech by human rights activists who are reporting having been targets of incitement to violence. To address this limitation, this thesis suggests further research regarding the possibility for platforms to create specific accounts with different settings that could be certified as protecting human rights activists to safely report cases of incitement to violence. These specific settings would need to be discussed with the human rights civil society and non-governmental organizations representatives to ensure its efficacy. Understandably, the creation of certified accounts for human rights activists may put them at a higher risk. Nevertheless, future studies are needed to explore how to best protect such human rights activists' accounts. For example, research could explore the possibility to protect these accounts by purposefully mislabelling them, *i.e.* purposefully miscategorising such accounts as normal users and not as human rights activists, and by applying several layers of encryption.

The third specific area for further research identified in Chapter 4 concerns the need to study the human rights responsibilities for platforms providing E2EE services beyond messaging applications as online platforms increasingly deploy innovative E2EE services. For example, monetization features on E2EE services, *i.e.* enabling easy purchases and selling functionalities, can facilitate the access to illegal goods such as weapons and increase the risk of human rights violations.

58 Karen Hao, MIT Technology Review, Artificial Intelligence, We read the paper that forced Timnit Gebru out of Google. Here's what it says. (2020), available at <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/> accessed 26 August 2024.

The specific areas requiring further research which were identified in Chapter 5 relate to the need to examine whether the operationalization of the DSA-established out-of-court mechanism complies with the legitimacy, promptness, and impartiality criteria essential for remedial processes. Regarding the remedial outcomes, further studies are needed to monitor technological developments on the platforms functionalities with the goal to explore new human rights compliant business models and the employment of new algorithmic decisions on content management to remediate people harmed by hate speech amplified by the platforms. Regulation should follow and direct such digital advancements in the field of content and platform regulation.