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

Nave, E.V.R.

5 Human rights responsibilities of online platforms to remediate criminal hate speech

A call for a thorough corporate remedial responsibilities framework in Europe for criminal hate speech attributable to online platforms[12]

ABSTRACT

Online platforms have adopted business models enabling the proliferation of hate speech. In some extreme cases, platforms are being investigated for employing algorithms that amplify criminal hate speech such as incitement to genocide. Legislators have developed binding legal frameworks clarifying the human rights due diligence and liability regimes of these platforms to identify and prevent hate speech. Some of the key legal instruments at the European Union level include the Digital Services Act, the proposed Corporate Sustainability Due Diligence Directive, and the Artificial Intelligence Act. 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. This Chapter addresses this legal vacuum by proposing 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.

5.1 INTRODUCTION

Business models adopted by online platforms[3] have contributed to the proliferation of online hate speech. Frances Haugen, a whistleblower from Meta Platforms, Inc. (formerly Facebook, Inc.) revealed that the platform prioritized

---

1 This Chapter is currently under review at a peer-reviewed scientific journal.

2 References to the following legal and policy frameworks were updated to reflect the latest available information: the Council of Europe Committee of Ministers Recommendation CM/Rec(2022)16; the European Union Regulation of the European Parliament and of the Council on a Single Market for Digital Services (DSA); the European Union Regulation of the European Parliament and of the Council Laying down harmonized rules on artificial intelligence (AI Act); the European Union Directive of the European Parliament and of the Council on combating violence against women and domestic violence; and, the European Union Directive of the European Parliament and of the Council on corporate sustainability due diligence (CSDDD). Cross-references should be read as referring to other references within the present Chapter.

3 Online platforms as per the DSA (also referred to as social media companies). This research employs businesses, companies interchangeably, and assumes that online platforms fall under these categories.

growth over countering online hate speech in countries such as Afghanistan, Ethiopia, and India.[4] In a more extreme example, Amnesty International and the United Nations alerted to Meta's significant contribution to the genocide of the Rohingya in Myanmar after its algorithms failed to take down and amplified hate speech towards this Muslim community.[5] Other online platforms have also been under increased scrutiny for adopting content moderation and recommendation algorithms amplifying hate speech.[6]

The framework addressing the companies' responsibilities to comply with human rights is thoroughly developed in the United Nations Guiding Principles on Businesses and Human Rights (UNGPs).[7] The UNGPs, though not legally binding, were endorsed by the United Nations Human Rights Council in 2011 and are the key international standard-setting instrument explaining the three essential corporate human rights responsibilities. Based on the UNGPs, companies must adopt: (i) a policy commitment to respect human rights; (ii) a human rights due diligence process to identify, prevent, and mitigate adverse impacts on human rights; and, (iii) remediation mechanisms of any adverse impacts on human rights that the company caused or contributed to.[8]

At the European Union (EU) level, online platforms have the corporate human rights responsibility to counter illegal content, including hate speech. The Corporate Sustainability Due Diligence Directive (CSDDD),[9] the Artificial

4 Isabel Debre and Fares Akram, 'Facebook's language gaps weaken screening of hate, terrorism' (2021) https://apnews.com/article/the-facebook-papers-language-moderation-problems-392cb2d065f81980713f37384d07e61f?utm_campaign=SocialFlow&utm_source=Twitter&utm_medium=AP(accessed 28 May 2024).

5 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); Human Rights Council, 'Report of the independent international fact-finding mission on Myanmar' (2018) A/HRC/39/64, https://www.ohchr.org/en/press-releases/2018/09/myanmar-un-fact-finding-mission-releases-its-full-account-massive-violations?LangID=E&NewsID=23575 (accessed 28 May 2024), Para. 74.

6 Rachel Griffin, 'The Law and Political Economy of Online Visibility. Market Justice in the Digital Services Act' Technology and Regulation 2023 (2023): 69-79. See also AlJazeera, 'The Listening Post: Genocide in Gaza: Enabled by AI, powered by Big Tech' (2024) available at <https://www.aljazeera.com/program/the-listening-post/2024/4/13/genocide-in-gaza-enabled-by-ai-powered-by-big-tech> accessed 30 May 2024.

7 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 (UNGPs).

8 UNGPs (note 7), Principle 15.

9 European Union, Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937 (CSDDD), available at <https://www.europarl.europa.eu/doceo/document/TA-9-2024-0329_EN.pdf> accessed 29 May 2024.

Intelligence Act (AI Act),[10] the Digital Services Act (DSA),[11] the Audiovisual Media Services Directive (AVMSD)[12] all contribute to establishing the human rights due diligence of online platforms to counter online hate speech. Nevertheless, this European legal framework fails to clarify the third task stemming from the UNGPs, i.e. the remedial responsibilities of online platforms to redress people harmed[13] by online hate speech caused or contributed to by the platforms.

This Chapter's central research question is two-fold: In compliance with the right to an effective remedy, how can European legislators better align the framework on corporate remedial responsibilities of online platforms which caused or contributed to criminal hate speech with the general framework on corporate remedial responsibilities? Additionally, are there heightened remediate responsibilities for very large online platforms (VLOPs)[14] or for cases of criminal hate speech amounting to gross violations of human rights?

This Chapter covers legal and policy instruments from both the European Union and the Council of Europe given the alignment between the two human rights systems.[15] Occasional references to international human rights instruments contextualize their influence on European instruments. Doctrinal research identifies legal loopholes in legislation and suggests normative approaches compliant with human rights. This Chapter focuses on hate speech on online platforms for two reasons. First, online platforms, and especially VLOPs, have constituted the most problematic digital environment quickly disseminating hate speech. Second, online platforms are increasingly regulated at the European level and thus allow for a more consolidated normative analysis.

To answer the research question, Section 5.2. analyses the European standards on criminal hate speech. Given that there is no definition of criminal

10 European Union, Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence and amending certain Union legislative acts COM(2021) 206 final (AI Act) available at <https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf> accessed 28 May 2024.

11 European Union, Regulation of the European Parliament and of the Council on a Single Market For Digital Services and amending Directive 2000/31/EC (DSA), Art. 93.

12 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 (AVMSD), OJ L 95.

13 This research recognizes the civil society arguments against legal expressions patronizing the agency of marginalized people and thus avoids the use of "victims" and "protected characteristics", and uses instead people targeted by hate speech.

14 DSA, note 11, Art. 41.

15 Steven Greer, Janneke Gerards, and Rose Slowe. Human rights in the Council of Europe and the European Union: achievements, trends and challenges' (2018).

hate speech at the EU level,[16] the central instrument investigated is CM/Rec(2022)16 adopted by the Council of Europe Committee of Ministers.[17] This section also initiates the academic debate about the elements of criminal hate speech that may classify as gross violations of human rights. In these cases, the international standards on the right to remedy for gross violations of human rights should apply.

Facebook's contribution to the genocide of the Rohingya in Myanmar is used as an example mainly in Section 5.2, but also occasionally referred to in other sections. This case is relevant because it is one of the most thoroughly documented showing the societal impact of the corporate human rights responsibilities of VLOPs contributing to hate speech as well as the impact of the lack of compliance with corporate remedial responsibilities.

Section 5.3. investigates the application of the right to effective remedy prescribed in Art. 13 of the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR),[18] Art. 47 of the Charter of Fundamental Rights of the European Union (CFREU),[19] and the Victims Rights Directive[20] to online hate speech. This section also examines the international standards on the right to remedy for cases of gross violations of human rights.

Section 5.4. clarifies the general corporate remedial responsibility by explaining the framework stemming from the UNGPs and from the Organisation for Economic Co-operation and Development Guidelines for Multinational Enterprises and OECD Due Diligence Guidance (OECD Guidelines).[21] This framework covers: modes of responsibility; remedial processes, and remedial outcomes. This framework applies to online platforms that caused or contributed to criminal hate speech.

Section 5.5. highlights the need for and proposes legal standards for a corporate remedial responsibilities framework at the EU level, including for online platforms that caused or contributed to criminal hate speech. The legal instruments reviewed are the CSDDD, the AIA, the DSA, the AVMSD, and the policy instruments researched are the Code of Conduct on countering

16 European Commission, 'No place for hate in Europe. Commission and High Representative launch call to action to unite against all forms of hatred' (2023) https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6329 (accessed 28 May 2024).

17 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).

18 Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950.

19 European Union, Charter of Fundamental Rights of the European Union (2007/C 303/01), C 303/1, 14 December 2007.

20 European Union: Council of the European Union, Directive 2012/29/EU of the European Parliament and of the Council of October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA, L 315/57, 14 November 2012.

21 OECD, 'OECD Guidelines for Multinational Enterprises' (2011); OECD, 'OECD Due Diligence Guidance for Responsible Business Conduct' (2018).

illegal hate speech online, and Recommendations CM/Rec(2022)16 and CM/Rec(2014)6.[22] The proposed standards focus on clarifying modes of responsibilities, remedial processes, and remedial outcomes. In this context, the three remedial outcomes analysed are guarantees of non-repetition, restitution, and compensation.

5.2 Criminal hate speech on online platforms

5.2.1 European standards on criminal hate speech

Although there is no binding definition of hate speech in international or European human rights law, CM/Rec(2022)16[23] distils the key elements for the regulation of hate speech both online and offline. CM/Rec(2022)16 clarifies that hate speech is always illegal as it is either (1) criminalized in its most severe forms, or (2) prohibited under civil or administrative law.[24]

This Chapter explores the legal framework applicable to category (1), i.e. criminal hate speech.[25] The decision to focus on criminal hate speech is based on a growing recognition of its key elements at the European level, specifically following the adoption of CM/Rec(2022)16.[26] CM/Rec(2022)16 clarifies, in Paragraph 11, the expressions that are criminally actionable based on existing international and regional human rights.[27]

CM/Rec(2022)16 takes an open-ended approach to the list of impermissible grounds[28] for hate speech as both Paragraph 11 and Paragraph 2 introduce

22 Council of Europe Committee of Ministers, Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a Guide to human rights for Internet users (CM/Rec(2014)6).

23 CM/Rec(2022)16, note 17.

24 CM/Rec(2022)16, note 17, Explanatory memo, Para. 54.

25 Hereinafter, this research employs "criminal hate speech" and "the most severe forms of hate speech" interchangeably.

26 Such increased understanding of the criminal hate speech allows for an extended legal reasoning on the States' positive obligations to protect people targeted by hate speech as well as on the corporate human rights responsibilities of online platforms required to counter online hate speech.

27 CM/Rec(2022)16, note 17, Para. 11. For a verbatim reading of Paragraph 11 of CM/Rec(2022)16, see Section 2.5.2.3. of this thesis.

28 Tarlach McGonagle 'Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas' (2011). 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 recognizes 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.

a list of several characteristics by using "such as".[29] Nevertheless, this Chapter defends that CM/Rec(2022)16 could have improved legal coherence had it expressly referred to two elements stemming from the critical legal conceptualization of hate speech. First, the historical oppression perpetuated by hate speech[30] and, second, the intersectionality of systems of oppression with a view to adequately reflect the harm caused by hate speech.[31] Hence, the subsequent analysis in this Chapter adopts an explicitly open-ended conceptualization of impermissible grounds for hate speech, grounded in the acknowledgement that hate speech is used to perpetuate systems of oppression, and that the intersectionality of historical systems of oppression is an aggravating factor harming people targeted by hate speech.

At the EU level, the European Commission published in 2021 a Communication encouraging the Council of the European Union (Council) to extend hate speech and hate crime to the list of EU crimes under Art. 83(1) of the Treaty on the Functioning of the European Union (TFEU).[32] However, whilst the EU does not adopt such legislation on criminal hate speech, this Chapter follows the conceptualization of criminal hate speech in Paragraph 11 CM/Rec(2022)16.

Finally, certain elements of criminal hate speech *may* classify as gross violations of human rights. Though there is no universally agreed definition of the term "gross violations of human rights",[33] a guiding reference providing a clearer conceptualization of the meaning of the term is Paragraph 30 of the 1993 of the United Nations Vienna Declaration and Program of Action: "Gross and systematic violations.. .include, as well as torture and cruel, inhuman and degrading treatment or punishment, summary and arbitrary executions, disappearances, arbitrary detentions, all forms of racism, racial discrimination and apartheid, foreign occupation and alien domination, xenophobia, poverty, hunger and other denials of economic, social and cultural rights, religious intolerance, terrorism, discrimination against women and lack

29 CM/Rec(2022)16, note 17, Paragraphs 2 and 11.

30 Katharine Gelber, 'Differentiating hate speech: a systemic discrimination approach' Critical Review of International Social and Political Philosophy (2019).

31 Eugenia Siapera and Paloma Viejo-Otero. "Governing hate: Facebook and digital racism." Television & New Media 22.2 (2021): 112-130.

32 European Commission, 'A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime' (2021) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021DC0777 (accessed 28 May 2024); Art. 83(1) of the TFEU specifies a list of areas of crime where the European Union legislators may establish minimum legal thresholds regarding the definition of criminal offences and sanctions applicable in all Member States of the EU.

33 See e.g., Roger-Claude Liwanga (2015) The Meaning of Gross Violation of Human Rights: A Focus on International Tribunals' Decisions over the DRC Conflicts, 44 Denv. J. Int'l L. & Pol'y 67, 69-73.

of the rule of law".[34] Nevertheless, the application of the concept in international and regional instruments has been inconsistent,[35] and its meaning remains debatable.

In this context, the normative framework presented in this Chapter cannot clarify which, if any, elements of criminal hate speech amount to gross violations of human rights. Even though Paragraph 11 of the CM/Rec(2022)16 includes, together with incitement to genocide, also incitement to crimes against humanity, and incitement to war crimes as criminal hate speech; it should also be noted that international criminal law does not clarify if these three types of incitement would classify as the most serious crimes in international law amounting to gross violations of human rights.[36]

Not pertaining to resolve this discussion, this Chapter seeks to acknowledge the possibility that elements of criminal hate speech may amount to gross violations of human rights and thus result in the application of the frameworks protecting the right to remedy and reparation for victims of gross human rights violations. This analysis is key to adequately frame the corporate remedial responsibilities of online platforms responsible for such criminal hate speech potentially amounting to gross violations of human rights.

5.2.2 The role of online platforms

This section introduces, first, the services provided by online platforms and, second, how they facilitate the spread of hate speech on their platforms. After that, this section expands on Meta's contribution to the genocide of the Rohingya in Myanmar as an example clarifying the problematic role of online platforms contributing to the rise of online and offline hate speech.

Online platforms facilitate the dissemination of user-generated content.[37] Given the large user base and high amounts of content, online platforms typically employ two types of algorithms to manage content:[38] (1) content

34 See World Conference on Human Rights, Vienna Declaration and Programme of Action, 1 30, U.N. Doe. A/CONF. 157/23 (June 25, 1993). See also Definition of Gross and Large-scale Violations of Human Rights as an International Crime, Comm. on Human Rights, Prevention of Discrimination and Protection of Minorities, Working paper submitted by Mr. Stanislav Chemichenko in accordance with Sub-Comm. decision 1992/109, 14, U.N. Doc. E/CN.4/Sub.2/1993/10 (June 8, 1993).

35 With legal instruments employing multiple terms, such "gross", "grave", "serious".

36 Art. 25(3)(e) of the Rome Statute criminalizes direct and public incitement of other to commit genocide. Mark Klamberg, ed. Commentary on the law of the International Criminal Court. Vol. 29. Torkel Opsahl Academic EPublisher, 2017. See Neema Hakim, "How social media companies could be complicit in incitement to genocide." Chi. J. Int'l L. 21 (2020): 83.

37 Michael Luca, 'User-generated content and social media' Handbook of media Economics. Vol. 1. North-Holland, 2015. 563-592.

38 Covering the management of both users' accounts and users' posts.

moderation algorithms, and (2) content ranking and recommendation algorithms.[39]

Content moderation algorithms are used to enforce policies of prohibited content. Users are informed about the content that is prohibited on the platform in the terms of service.[40] Examples of outcomes of content moderation include disabling, labelling, suspension, and removal of content.[41] terms of service often do not clarify the standards used to decide on content moderation outcomes. The current regulatory framework applicable to ToS provides insufficient guidance regarding the content that should be prohibited[42] or the way that ToS should address the outcomes to be attained from content moderation.[43]

Ranking and recommendation algorithms assist with the task of deciding which content to first display on the users' newsfeed or on auto-plays after the completion of a given video. The suggestion of subsequent content that is ranked high is called chaining.[44] The reverse operation, when a content is deliberately not suggested, is called demotion or down-ranking. These algorithms typically aim to link users to other users, groups, or to specific posts that can match their interests and thus maximize engagement on the platform.[45] Online platforms have disclosed little to no information on the internal processes guiding these ranking and recommendation algorithms or possible outcomes.[46]

---

39 Tarleton Gillespie, 'Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.' Yale University Press, 2018.

40 The platforms typically require that users agree to the terms of service when creating an account. This research refers to terms and conditions and community guidelines interchangeably.

41 Eric Goldman, 'Content moderation remedies' Mich. Tech. L. Rev. 28 (2021): 1., 24; Eline Labey and Valentina Golunova (2022). 'Judges of Online Legality: Towards Effective User Redress in the Digital Environment' In European Yearbook on Human Rights (1 ed., pp. 105-135). Intersentia.

42 João Pedro Quintais, Naomi Appelman, and Ronan Ó. Fathaigh. 'Using terms and conditions to apply fundamental rights to content moderation' German Law Journal 24.5 (2023): 881-911; Eva Nave and Lottie Lane, 'Countering online hate speech: How does human rights due diligence impact terms of service?' Computer Law & Security Review 51 (2023): 105884.

43 E.g., CM/Rec(2022)16 Paragraph 23 recommends that Member States regulate the necessity that internet intermediaries explain a decision to block, take down, or deprioritize certain content. However, it could have provided more detailed guidance for content moderation had it clarified the suitability of moderation outcomes depending on the severity of hate speech.

44 Tarleton, note 39.

45 Paddy Leerssen, 'An End to Shadow Banning? Transparency rights in the Digital Services Act between content moderation and curation' Computer Law & Security Review 48 (2023): 105790.

46 Some online platforms have created dedicated websites to explaining their content moderation practices. E.g., https://transparency.x.com/en.html for twitter, https://transparency.fb.com/en-gb/ for Meta, https://about.linkedin.com/transparency for LinkedIn (accessed 28 May 2024).

The Committee of Ministers of the Council of Europe and the European Commission have warned that the algorithms employed by online platforms can facilitate the dissemination of online hate speech.[47] Analysing to what extent online platforms enhance the severity of hate speech, it is relevant to review the context in which the expression was manifested. When assessing the severity of hate speech, the ECtHR evaluates "contextual variables"[48] such as: the political and social context at the time of the speech;[49] the speaker's status or role in society;[50] the reach and form of dissemination of the speech;[51] the likelihood and imminence that the speech results, directly or indirectly, in harmful consequences;[52] the nature and size of the audience;[53] the perspective of the people targeted by the speech (including its historical oppression).[54]

This Chapter explores how online platforms affect the severity of hate speech by reviewing three contextual variables: (1) reach, as well as the size of the audience; (2) the polarized and susceptible nature of the audience; and, (3) the likelihood of harm. These three variables were selected based on the algorithms currently discussed within the context of online platforms.

First, online platforms typically enable faster dissemination of content to larger audiences than traditional offline media, thereby amplifying the reach of speech. Users can instantaneously publish content with a wider network than in offline settings. Nevertheless, studies show that the reach of speech is only increased for certain types of content *e.g.*, hate speech spreading faster than innocuous content.[55] Depending on the algorithms deployed, content can be amplified, deamplified, blocked, removed, etc. Typically, algorithms

47 CM/Rec(2022)16, note 17, Preamble and Explanatory memo, Para. 86; European Commission, note 32.

48 Michel Rosenfeld, 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 (2002). CM/Rec(2022)16, , note 17, Explanatory memo, Para. 32.

49 Leroy v. France Para. 38; Delfi AS v. Estonia paras. 142-146; Perinçek v. Switzerland Para. 205.

50 Féret v. Belgium, no. 15615/07, 16 July 2009, Para. 63; General recommendation No. 35, Combating racist hate speech of the Committee on the Elimination of Racial Discrimination; the Rabat Plan of Action on the prohibition of advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence; and, the Guide on Article 10 of the ECHR, Freedom of expression, Para. 225.

51 Savva Terentyev v. Russia, no. 10692/09, Para. 79; Delfi AS v. Estonia Para. 110; Stomakhin v. Russia, no. 52273/07, 9 May 2018, Para. 131; and, Jersild v. Denmark paras. 32-33.

52 Perinçek v. Switzerland Para. 205; Savva Terentyev v. Russia paras. 32-33.

53 Vejdeland and Others v. Sweden, no. 1813/07, 9 February 2012, paras. 51-58; and Lilliendahl v. Iceland, no. 29297/18, 11 June 2020, paras. 38-39.

54 Budinova and Chaprazov v. Bulgaria Para. 63.

55 Binny Mathew et al., 'Spread of Hate Speech in Online Social Media' (2019) Proceedings of the 10th ACM Conference on Web Science 173 (accessed 28 May 2024).

are not trained to process either the context or the languages of already marginalized communities, resulting in the illegal removal of content produced by these communities.[56] Additionally, it is widely reported that platforms have been prioritizing user engagement often at the expense of human rights, such as the prohibition of discrimination.[57] For example, the Facebook Papers[58] revealed that ranking and recommending algorithms prioritized virality of content, often disregarding whether content is harmful or incites to violence.[59] Consequently, online platforms have increased the reach of hate speech.

Second, online platforms polarize large audiences of users due to their content recommendations algorithms. Designed to connect like-minded people, online platforms have facilitated the organization of "hate mongers",[60] and enabled offline violence.[61] In fact, the Wall Street Journal found that, in 2016, 64% of new members in extremist groups on Facebook in Germany resulted from algorithm recommendations.[62]

Third, by amplifying online hate speech and by polarizing users, the current algorithms increase the likelihood of harm. Amnesty International has explained how Meta's content moderation algorithms failed to take down content advocating for hatred, discrimination, and genocide of the Rohingya Muslim community in Myanmar.[63] This hateful content was then amplified

56 Janice Asare, 'Are Marginalized Communities Being Censored Online' (2020) Forbes https://www.forbes.com/sites/janicegassam/2020/05/24/are-marginalized-communities-being-censored-online/ (accessed 28 May 2024). Furthermore, online platforms often outsource the traumatic human review in content moderation to already marginalized communities working under extremely precarious work conditions; e.g., Adrienne Williams, Milagros Miceli and Timnit Gebru 'The Exploited Labor Behing Artificial Intelligence' (2022) https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/ (accessed 28 May 2024).

57 Larry Elliot, 'Big tech firm recklessly pursuing profits from AI, says UN head' (2024) The Guardian, https://www.theguardian.com/business/2024/jan/17/big-tech-firms-ai-un-antonio-guterres-davos (accessed 28 May 2024); Alyan Layug, et al. 'The impacts of social media use and online racial discrimination on Asian American mental health: cross-sectional survey in the United States during COVID-19." JMIR formative research 6.9 (2022): e38589 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9488547/ (accessed 28 May 2024); Allyson M Ganster 'Black women and digital resistance: The impact of social media on racial justice activism in Brazil and the United States' Diss. 2019, https://repositories.lib.utexas.edu/items/45168a42-b43d-47ea-800f-24cd7d2d04cc (accessed 28 May 2024).

58 A Wall Street journal investigation resulting from the work of former Facebook employee and whistle blower Frances Haugen.

59 Amnesty International, note 5, 42.

60 Damon Henderson Taylor, 'Civil Litigation against Hate Groups Hitting the Wallets of the Nation's Hate-Mongers' Buff. Pub. Int. LJ 18 (1999): 95.

61 Amnesty International, note 5, 42.

62 Amnesty International, note 5, 44; The Wall Street Journal, 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive' (2020), wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499 (accessed 28 May 2024).

63 Amnesty International, note 5.

by their ranking algorithm designed to maximise the users' engagement by showing such content at the top of newsfeeds. Moreover, hateful videos were also amplified by Facebook when its recommendation algorithm automatically played them in its "Up Next" feature. The United Nations Independent International Fact-Finding Mission on Myanmar concluded that "[t]he role of social media [was] significant" in the atrocities.[64]

The Rohingya are seeking remediation from Meta in three judicial actions, including a request for a USD $1 million for an educational project in the refugee camps. Despite admitting to not have done enough to prevent the platform from being used to incite offline violence,[65] Meta refuses to remediate through the educational project, communicating that it had instead improved its content moderation algorithms.[66] Meta does not detail in which way it has improved its algorithms and Amnesty International emphasizes compliance with remediation responsibilities must address the victims' harms.[67]

5.3 Right to remedy for criminal hate speech online

Having clarified the conceptualization of criminal hate speech employed in this Chapter,[68] this section explains the operationalization of the human right to an effective remedy of people targeted by criminal hate speech. This section identifies, first, the harm caused by criminal hate speech including on online platforms (Section 5.3.1), then sets out the European standards on the State's duty to ensure access to an effective remedy for people targeted by criminal hate speech (Section 5.3.2).

5.3.1 Harm caused by hate speech

Critical race theory was the legal scholarship to first advance the conceptualization of harms caused by hate speech.[69] According to this scholarship, hate speech can cause psychological, physical, and economic or material harms.[70] Critical race scholars also stressed the cumulative effect of continued exposure to hate speech.[71]

64 United Nations, note 5.
65 United Nations, note 5, Para. 74.
66 Amnesty International, note 5.
67 Amnesty International, note 5.
68 Section 2.
69 Richard Delgado 'Understanding words that wound'. Routledge, 2019.
70 Eva Nave, 'Hate Speech, Historical Oppressions, and European Human Rights' Buff. Hum. Rts. L. Rev. 29 (2022): 83, 91.
71 Richard Delgado, 'Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling' (1982) 17 Harvard Civil Rights Liberties Law Review 133.

The psychological harms experienced by people targeted by hate speech range from fear, anger, low self-esteem, low capacity of attention, withdrawal from society, depression, nightmares, post-traumatic stress, psychosis.[72] Studies show that these harms have an aggravated impact on younger people and children.[73] These layers of harm passed through generations lead to an increased difficulty in dealing with the psychological harms caused by hate speech.[74] Furthermore, access to psychological support is limited because it is not just expensive but also practitioners often come from privileged backgrounds and thus lack the lived experience of people historically targeted by hate speech.[75]

The physical harms that people targeted by hate speech face can be distinguished between short-term and long-term physical harms. Short-term physical harms include accelerated breathing and heart rate, dizziness, headaches, and raised blood pressure.[76] In the most serious cases, hate speech inciting to violence can lead to hate crimes, war crimes, genocide, or crimes against humanity.

Hate speech may also cause economic or material harms of the people it targets. Hate speech may jeopardize access to e.g., education, health, or employment, if by continued exposure to hate speech, people are forced to leave their studies, jobs, neighbourhoods, cities, or countries, or to avoid public spaces altogether. In some of the most extreme cases, people targeted by hate speech may become refugees seeking asylum, often facing dire situations ranging from insecurity to lack of access to water and other basic human rights.

In the specific context of harms experienced by people targeted by criminal hate speech on online platforms, all of the harms mentioned above apply i.e. psychological, physical, and economic harms. Additional impacts to consider include e.g., disengaging from online platforms to avoid exposure to hate speech may limit the exercise of access to information and freedom of assembly or association.[77]

72 Richard Delgado, note 71.
73 Joe R. Feagin and Debra Van Ausdale 'The first R: How children learn race and racism' Rowman & Littlefield Publishers, 2001.
74 Richard Delgado, note 71.
75 Gene Combs 'White privilege: what's a family therapist to do?' Journal of marital and family therapy 45.1 (2019): 61-75.
76 Richard Delgado, note 71; Research indicates that a potential cause for the higher number of deaths of African Americans associated with hypertension may be linked to continued exposure to hate speech.
77 Katharine Gelber, note 30.

5.3.2 State's duty to ensure access to remedy

5.3.2.1 European standards on remedies

People harmed by hate speech (whether online or offline), and especially by criminal hate speech, have the right to an effective remedy. The right to an effective remedy is a fundamental human right under international and European human rights law.[78] This right derives from a general legal principle that every breach on international law results in an obligation to provide remedy.[79] This Chapter focuses primarily on the European standards.

At the Council of Europe level, Art. 13 of the ECHR establishes the right to an effective remedy before a national authority. This provision lays down the State's positive obligation to investigate allegations of violations, including by private companies, of human rights in a "diligent, thorough, and effective" manner.[80] The national authority may be a judicial or non-judicial body, if the latter fulfils the independence and impartiality prerequisites.[81] It is essential that remedies are "available, known, accessible, affordable, and capable of providing adequate redress".[82] Importantly, the national authorities have the primary responsibility to investigate violations of human rights and a person may only appeal to the ECtHR after exhausting all available domestic procedures.

The right to remedy exists when there is an "arguable" grievance under the ECHR.[83] This means that Art. 13 of the ECHR is complementary to other rights[84] and may be invoked in two circumstances. First, if there is an allegation of a violation of another right in the ECHR. Second, if the person cannot effectively exercise the right to remedy at the national level.[85]

Finally, according to Art. 13 of the ECHR, the remedy must directly remediate the violation.[86] Nonetheless, in light of the margin of appreciation

78 Wojciech Piątek, 'The right to an effective remedy in European law: significance, content and interaction' China-EU Law Journal 6.3-4 (2019): 163-174.

79 Kathleen Gutman, 'The Essence of the Fundamental Right to an Effective Remedy and to a Fair Trial in the Case-Law of the Court of Justice of the European Union: The Best Is Yet to Come?' German Law Journal 20.6 (2019): 884-903.

80 Council of Europe, Effective Remedies Explanatory Memorandum, https://www.coe.int/en/web/freedom-expression/effective-remedies-explanatory-memo (accessed 28 May 2024).

81 Council of Europe, Guide on Article 13 of the ECHR Right to an effective remedy, https://www.echr.coe.int/documents/d/echr/guide_art_13_eng (accessed 28 May 2024), paras. 3, 24, and 26.

82 Council of Europe, Effective Remedies, https://www.coe.int/en/web/freedom-expression/effective-remedies#:~:text=You%20have%20the%20right%20to,pursue%20legal%20action%20straight%20away (accessed 28 May 2024).

83 Council of Europe, note 81, Para. 10.

84 Council of Europe, note 81, Para. 11.

85 Council of Europe, note 81, https://www.echr.coe.int/documents/d/echr/guide_art_13_engPara. 20.

86 Pine Valley Developments Ltd and Others v. Ireland, Commission decision, 1989.

afforded to Contracting States,[87] there is no specific prescription of the adequate form of remedy.[88] Instead, the effectiveness of the remedy should be evaluated on a case-by-case basis.[89]

At the European Union level, Art. 47 of the CFREU prescribes that "Everyone whose rights and freedoms guaranteed by the law of the Union are violated has the right to an effective remedy before a tribunal in compliance with the conditions laid down in this Article (…)".[90] While the provisions in the CFREU with corresponding rights in the ECHR must be interpreted with similar meaning and scope to the provisions in the ECHR, there is a key difference between Art. 13 of the ECHR and Art. 47 of the CFREU. Art. 47 of the CFREU stipulates that the competent national authority must be a judicial institution. This may be interpreted as strengthening the right since judicial bodies will in principle by default be independent and impartial, while other non-judicial bodies may not be. Notwithstanding, this requirement may also place an added burden on the judicial system and may result in more constraints to exercise the right to an effective remedy.

Additionally, crime survivors in the EU are covered by the Victims' Rights Directive which establishes minimum requirements for rights, assistance, and protection of crime survivors.[91] Key rights include the right to legal aid such as the right to a fair remedy,[92] the right to return of property, and the right to compensation.[93] Whilst the EU does not include hate speech in the EU list of crimes, the Victims' Rights Directive applies only to elements of hate speech criminalized in the EU.[94]

Applying the European framework on the right to effective remedy established by the CoE and by the EU to cases of online hate speech, two remarks are due. First, it is clear that national authorities have the duty to protect, investigate, and ensure access to remedies. This framework applies to acts committed in digital settings by either users or internet intermediaries *e.g.*, criminal hate speech.[95] Importantly, remedial avenues must be available, known, accessible, and affordable.

87 Budayeva and Others v. Russia, 2008, Para. 190.
88 Council of Europe, note 81.
89 Colozza and Rubinat v. Italy, Commission decision, 1982, 146-147.
90 CFREU, note 19, Art. 47.
91 European Union, Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA (Victims Directive).
92 European Commission, DG Justice Guidance Document related to the transposition and implementation of the Victims Directive, 34, https://commission.europa.eu/document/download/238cafb6-d5cd-4d1a-8624-a0bafb2cdfa3_en?filename=13_12_19_3763804_guidance_victims_rights_directive_eu_en.pdf (accessed 29 May 2024).
93 Victims Directive, note 91, Arts. 15 and 16.
94 Victims Directive, note 91, Art. 1.
95 Council of Europe, note 82.

Second, there are different legal thresholds at both the CoE and the EU level regarding the competent authority with which to lodge a remedy claim. Given the extensive work on the right to remedy developed by the Council of Europe for cases of criminal acts online and also recognizing that effective processes may at times be found outside judicial settings, this Chapter follows the approach that remedies can be sough with both judicial and non-judicial institutions, as long as these are independent and impartial.

5.3.2.2 Remedies for gross human rights violations

As mentioned in Section 5.2.1, some elements of criminal hate speech may amount to gross human rights violations. In these cases, the international and the European frameworks on the right to remedy and reparation for victims of gross violations of human rights law are complementary and should apply.

At the international level, States are obliged to: (a) prevent violations; (b) effectively, promptly, thoroughly, and impartially investigate violations and, when necessary, take action against those responsible; (c) provide alleged victims with equal and effective access to justice; and, (d) provide effective remedies.[96] This framework calls for States to adopt provisions for universal jurisdiction.[97] Importantly, the conceptualization of victims includes persons individually or collectively harmed physically, psychologically, emotionally, economically, or who suffered substantial impairment of their fundamental rights.[98]

At the European level, the Council Decision enabling targeted restrictive measures to address serious human rights abuses worldwide applies.[99] The understanding of human rights abuses in this framework accounts for genocide and crimes against humanity, and extends to other human rights abuses if widespread and systematic.[100] The sanctions apply to both natural and legal persons, as companies.[101] For these natural or legal individuals, sanctions include *inter alia* asset freeze and a prohibition to make funds or economic resources available. Remarkably, this Council Decision establishes a global human rights sanctions regime providing the EU with a framework to target

96 United Nations, Resolution adopted by the General Assembly on 16 December 2005, 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, II(3).

97 United Nations, General Assembly Resolution 60/147, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violation of International Human Rights Law and Serious Violations of International Humanitarian Law, para 5.

98 A/RES/60/147, note 96, V(8).

99 European Union, Council Decision (CFSP) 2020/1999 of December 2020 concerning restrictive measures against serious human rights violations and abuses.

100 A/RES/60/147, note 96, Art. 1(1)(d).

101 CFSP, noted 99, Arts. 2 and 3(1).

inter alia companies responsible for serious human rights violations, regardless of where these took place.

Applying these regimes to survivors of criminal hate speech amounting to gross human rights violations, it becomes clear that States are obliged to ensure access to an effective remedy, including when harm was caused by businesses. Moreover, the conceptualization of survivor should include people directly and indirectly affected by the crime. Finally, businesses may be considered the perpetrators and thus may have to comply with restrictive sanctions, e.g., asset freeze measures. For example, the EU sanctions regime enables the EU to impose sanctions to Meta for its significant contribution to the genocide of the Rohingya in Myanmar.

The UN and EU standards on the right to an effective remedy for survivors of gross human rights violations offer clearer and more inclusive definitions of survivors, perpetrators, and remedial processes, than the general European standards on the right to an effective remedy (Section 5.3.2.1). First, while the general standards consider survivors only those directly impacted by the crime, the specific standards clarify that, for cases of gross violations of human rights, survivors are those affected both directly and indirectly. Second, the specific standards for victims of gross violations of human rights expressly foresee that non-state actors can be responsible. Third, the specific standards go beyond the general standards by explicitly calling States to implement universal jurisdiction and restrictive measures to address gross violations of human rights, including when committed by companies outside their territory. Applying these standards to criminal hate speech, follows that the EU legislators have a heightened duty to align corporate remedial responsibilities with the right to an effective remedy for criminal hate speech cases amounting to gross human rights violations.

5.4 General framework: corporate remedial responsibilities for online platforms

This section investigates the general remedial responsibilities when the harm is attributable to businesses, including online platforms, and clarifies the modes of corporate responsibility (Section 5.4.1), the remedial processes (Section 5.4.2), and the remedial outcomes (Section 5.4.3).

5.4.1 Modes of corporate responsibility

The UNGPs articulate corporate remedial responsibilities for businesses which caused or contributed to adverse impacts on human rights.[102] Adverse impacts on human rights happen when the exercise of said human right is excluded or reduced, and can be either actual or potential adverse impacts.[103] Actual impacts refer to an adverse impact that already occurred or is occurring, and potential impact refers to impact that has not occurred yet. Potential adverse impact can either be avoidable or unavoidable, the latter ultimately materializing as an actual adverse impact.

The general framework on corporate human rights remedial responsibility prescribes two modes of remedial responsibilities: the responsibility to remediate and the responsibility to use leverage.[104] The corporate responsibility to remediate, is encapsulated in Guiding Principle 22 of the UNGPs as follows:

> "Where business enterprises identify that they have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes".[105]

The OECD Guidance clarifies that this Principle 22 establishes the corporate responsibility to remediate actual adverse impacts that the company *caused*, *contributed to*, or potential but unavoidable adverse human rights impacts that the company will cause or contribute to. A business *caused* an actual adverse human rights impact when its operations alone resulted in the adverse impact.[106]

Conversely, a business is said to have *contributed to* an actual adverse impact on human rights when i) its operations, together with operations of other businesses, or ii) its operations alone, caused, facilitated or incentivized another business to cause an adverse impact on human rights. Notably, the contribution must be substantial.[107]

The second mode of corporate remedial responsibility encompasses the use of leverage to prevent or mitigate actual adverse impacts that the company was directly linked to, and for potential adverse impacts that are avoidable. A company is directly linked to an actual adverse human rights impact if the connection is not sufficiently substantial to amount to contribution. In these cases, the company is not required to remediate, but rather to use its leverage to influence the other actor causing the adverse effects to prevent or reduce

---

102 United Nations Human Rights, Office of the High Commissioner, Implementing the UN "Protect, Respect and Remedy Framework" (UNGPs Guide), 7.
103 UNGPs Guide, note 102, 5.
104 UNGPs Guide, note 102, Paras. 19, 21.
105 UNGPs, note 7, Principle 22.
106 OECD Due Diligence Guidance, note 21.
107 OECD Due Diligence Guidance, note 21, 70.

said negative effects.[108] Figure 5 summarizes the general framework on corporate remedial responsibilities.

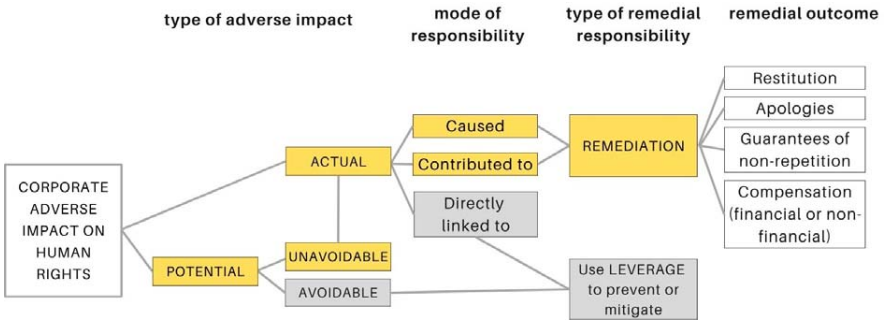

Figure 5 – Corporate remedial responsibilities for adverse human rights impacts

This general framework articulates remedial responsibilities for all businesses, including online platforms. The following sections investigate the remedial processes and outcomes of the corporate responsibility to remediate actual or unavoidable adverse impacts on human rights, including criminal hate speech caused or contributed to by online platforms.

5.4.2 Remedial processes

Remedial processes are the processes through which a remedial responsibility is assessed, and may either be ad-hoc or pre-established for specific adverse human rights impacts.[109] For businesses whose operations pose a high risk to human rights, a proactive approach in investigating their actual or potential adverse impact on human rights is advisable. In these cases, businesses should adopt an operational-level grievance mechanism[110] to enable individuals directly affected by the business' operations, to formally lodge concerns, complaints, and seek remedies.

Businesses may provide for remediation directly or in cooperation with another legitimate process.[111] Subsequently, there is no need for a prior judicial decision,[112] and businesses that acknowledge having caused or contributed to actual or unavoidable adverse human rights impacts have the responsibility to remediate. Nevertheless, when businesses do not provide remediation

108 OECD Due Diligence Guidance, note 21, 72.
109 UNGPs Guide, note 102, 70.
110 UNGPs, note 7, Principle 29.
111 UNGPs Guide, note 102, Q. 66.
112 UNGPs Guide, note 102, Q. 64.

proactively, State-based legitimate remedial processes should be initiated and businesses must collaborate.[113]

Applying these standards to online platforms, the functionality allowing users to report content arguably qualifies as an operational-level grievance mechanism. Nevertheless, this functionality alone does not fulfil the legitimacy criteria of remedial processes if not overseen by impartial bodies.[114] Additionally, the reporting process normally assesses whether content complies with terms of service and not with human rights standards.[115] For cases where the online platforms caused or contributed to criminal hate speech, if platforms do not comply with remedial processes, these should be initiated by States.[116] The standards on the individual right to remedy apply and, equally, the special regime on remedies for gross human rights violations applies to cases of criminal hate speech amounting to gross human rights violations.

5.4.3 Remedial outcomes

To determine the most appropriate remedial outcomes, businesses should seek to clarify what remedy the victims find most effective.[117] The general framework for remedial outcomes includes: restitution; satisfaction; rehabilitation; compensation; guarantees of non-repetition of harm.[118] These remedial outcomes were endorsed by the United Nations framework for cases of gross violations of human rights.[119]

These remedial outcomes apply to any businesses as online platforms which caused or contributed to criminal hate speech, including that amounting to gross human rights violations. Explaining in more detail what these outcomes entail, restitution aims to restore the original exercise of human rights before the violation and involves: restoration of liberty, identity, family life, and citizenship; return to the place of residence; restoration of employment; and, return of property.[120]

---

113 UNGPs Guide, note 102, Q. 66.

114 Kate Klonick, 'The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression.' Yale LJ 129 (2019): 2418; Rachel Griffin 'Rethinking rights in social media governance: human rights, ideology and inequality' European Law Open 2.1 (2023): 30-56.

115 Eva Nave and Lottie Lane, note 42.

116 United Nations, note 97, VII, Art.11(b).

117 UNGPs, note 7, Principle 20.

118 UNGPs Guide, note 102, Q. 64; United Nations, note 97, IX; Victor Stoica, 'Remedies before the International Court of Justice' Cambridge University Press, 2021.

119 A/RES/60/147, note 96.

120 A/RES/60/147, note 96, Para. 19.

Satisfaction aims to recognize the illegal acts that resulted in human rights violations and can be both pecuniary and non-pecuniary.[121] Some examples of satisfaction encompass: ceasing violations; verifying and publicly disclosing the facts (if not contributing to double victimization); searching of the disappeared or killed (in alignment with the victims' wishes); an official declaration or judicial decision restoring the victim's dignity, reputation and rights; judicial and administrative sanctions against those liable; tributes to the victims; inclusion of violations in training and educational material.

Rehabilitation aims to ensure the access to legal, medical and social services, including psychological support.[122] Compensation, similarly to satisfaction, can also be pecuniary and non-pecuniary and aims to repair any economically quantifiable harm. Such harm encompasses: physical or mental harm; lost opportunities, including as employment, education and social benefits; material damages and loss of earnings, including potential earnings; moral damages; costs deriving from legal, medical and social services, including psychological services.[123]

Finally, guarantees of non-repetition of harm should include: protecting human rights defenders; providing, on a priority and continued basis, human rights education; ensuring the observance of internal codes of conduct; promoting mechanisms for preventing and monitoring social conflicts; reviewing and reforming terms of service contributing to or allowing gross human rights violations.[124]

5.5 European framework: online platforms remedial responsibilities for criminal hate speech

This section examines the challenges with the current European framework on remedial responsibilities of online platforms which caused or contributed to criminal hate speech, including gross human rights violations (Section 5.5.1). After that, this section proposes standards to clarify and strengthen this framework by exploring the modes of responsibility, remedial processes, and three remedial outcomes (Section 5.5.2).

5.5.1 Challenges with current framework

This section studies the general framework on corporate remedial responsibilities in the EU CSDDD and AI Act (Section 5.5.1.1), the remedial responsibilities

121 Stoica, note 118, 146; A/RES/60/147, note 96, Para. 22.
122 A/RES/60/147, note 96, Para. 21.
123 A/RES/60/147, note 96, Para. 20.
124 A/RES/60/147, note 96, Para. 23.

in the DSA (Section 5.5.1.2), and the remedial responsibilities of online platforms in hate speech European sector-specific instruments (Section 5.5.1.3).

5.5.1.1 Corporate remedial responsibilities in the EU

The general legal framework on corporate remedial responsibilities in the EU stems from two instruments *i.e.*, the Corporate Sustainability Due Diligence Directive (CSDDD) and the Artificial Intelligence Act (AI Act). This framework applies to online platforms as these employ AI algorithms for content moderation.

The CSDDD seeks to ensure that businesses respect human rights within their operations and supply chains.[125] To achieve this goal, the CSDDD builds on the corporate human rights responsibilities framework established in the UNGPs and the OECD Guidelines, restating the corporate responsibilities to *inter alia* provide remedial mechanisms for human rights and environmental negative impacts caused by their operations, their subsidiaries and their value chains.[126] Nevertheless, the latest text of the CSDDD apparently fails to reflect the UNGPs' specific standards on remedial processes (*i.e.*, the importance of creating operational-level grievance mechanisms and the creation of adequate, legitimate, and impartial remedial processes) and on remedial outcomes (*i.e.*, restitution, satisfaction, compensation, rehabilitation, guarantees non-repetition). The CSDDD allows Member States the discretion to decide the means to reach the binding goals that it prescribes. As a result, in transposing this directive domestically, there may be States deciding to fully develop the corporate remedial responsibilities in alignment with the UNGPs.

The AI Act prescribes legally binding means to ensure that AI systems respect EU fundamental rights, while fostering investment and innovation.[127] The AI Act reflects the UNGPs and CSDDD overall standard on corporate human rights remedial responsibilities in two ways. First, it explains which AI systems do not comply with fundamental rights and are, therefore, prohibited. Art. 5 of the AI Act prohibits AI systems that deploy subliminal techniques capable of distorting a person's behaviour in a manner that causes or is likely to cause physical or psychological harm.[128] Applying this provision to online platforms, online platforms are undoubtedly prohibited from employing algorithms that amplify hate speech. Second, the AI Act prescribes a fundamental rights risk assessment framework to evaluate potential risks caused by AI systems.[129] This risk assessment aligns with the UNGPs corpor-

125 CSDDD, note 9, Recitals 6, 15, 25, 47, Art.3.
126 CSDDD, note 9, Recital 58.
127 AI Act, note 10, Recital 1.
128 AI Act, note 10, Art. 5. Notably, this standard seems to contradict the DSA no general monitoring requirement in Art. 7 because it requires platforms to monitor the impact of their algorithms and ensure that these are not enhancing the probability of harm.
129 AI Act, note 10, Recital 34.

ate human rights due diligence and remedial processes which require businesses to adopt processes to identify potential adverse human rights impacts.[130] Nevertheless, though expanding more than the CSDDD on the risk assessment, similarly to the CSDDD, the AI Act does not prescribe a comprehensive corporate remedial framework encompassing standards on remedial processes and outcomes to be achieved.

5.5.1.2 Remedial responsibilities in the Digital Services Act

The Digital Services (DSA) seeks to prevent illegal and harmful content online by regulating the human rights responsibilities[131] and liability[132] regimes of internet intermediary services operating within the EU. The conceptualization of internet intermediaries includes online platforms[133] *i.e.*, hosting services which store and disseminate to the public information produced by its users.[134]

The DSA prescribes different human rights responsibilities depending on the business' role, size, and impact.[135] Within the category of online platforms, the DSA attributes heightened human rights responsibilities to very large online platforms (VLOPs) *i.e.*, those with 45 million or more EU users per month.[136] In this context, VLOPs should identify, assess, and mitigate systemic risks, and negative effects for the exercise of fundamental rights.[137] Notably, hate speech is explicitly referred to as a systemic risk classified as illegal content in the EU.[138]

Reviewing the DSA framework on corporate remedial responsibilities, it is possible to conclude that the DSA does not provide a comprehensive approach to corporate modes of responsibilities, remedial processes, or remedial outcomes.

Firstly, the DSA does not clearly reflect the general UNGPs standards on the modes of corporate remedial responsibilities. Although Chapter II of the DSA regulates the liability regimes of internet intermediaries, it does not clarify that online platforms causing or contributing to adverse human rights impacts bear remedial responsibilities in line with the corporate responsibility frame-

130 UNGPs, note 7, Principle 15(b).
131 DSA, note 11, Chapter III.
132 DSA, note 11, Chapter II.
133 DSA, note 11, Recital 36.
134 DSA, note 11, Art. 2(h).
135 DSA, note 11, Art 33.
136 European Commission, Questions and answers on the Digital Services Act (2024) https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 (accessed 29 May 2024).
137 DSA, note 11, Art. 34 and 35.
138 DSA, note 11, Art. 34 (1)(a) and Recital 16.

work articulated in the UNGPs.[139] In another example, Art. 36 of the DSA prescribes that VLOPs must comply with specific crisis response measures in times of extraordinary serious threats to public security or public health in the EU, with the purpose of preventing, eliminating, or limiting said serious threats.[140] While this wording could be interpreted to reflect Principle 22 of the UNGPs, this link is not expressly mentioned. Moreover, Art. 36 of the DSA seems to apply only to VLOPs and in times of crisis, disregarding the ongoing nature of remedial responsibilities of all businesses regardless of size or crisis context.

Secondly, the DSA does not clearly expand on the general UNGPs standards on remedial processes. To clarify, the DSA refers to remedy as: i) the right to seek judicial remedies;[141] ii) an interim non judicial measure to ensure effective investigation of infringements, enforcement, or to prevent future infringements;[142] and, iii) an out-of-court dispute settlement for human rights infringements.[143] These elements seem to broadly reflect, respectively: i) the State's obligation to ensure the right to an effective remedy; ii) an operational-level grievance mechanism; and, iii) the legitimacy requirement for a non-judicial remedial process. However, these mechanisms require effective, impartial, and legitimate implementation and oversight. For example, concerns arise as to whether an out-of-court mechanism not empowered to impose binding decisions will provide access to an effective remedy.[144]

Thirdly, the DSA does not address the general UNGPs standards on remedial outcomes. In this context, the DSA missed an opportunity to provide harmonized guidance and steer this discussion on of best suited remedial outcomes for online harms caused or contributed to online platforms, including (criminal but not limited to) hate speech.

As a result, the DSA does not articulate a solid or comprehensive framework on the corporate human rights remedial responsibilities of internet intermediaries, including online platforms. This Chapter defends that, similarly to the UNGPs, remedial responsibilities, processes, and outcomes ought to have been addressed in the DSA as a whole and all together either under Chapter II after the liability provisions, or independently in a separate chapter on remedial responsibilities.

---

139 The due diligence obligations in Chapter III of the DSA can however be interpreted as creating a general duty of care which, if infringed would lead to liability. Machado CCV, Aguiar TH. Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models. Business and Human Rights Journal. 2023;8(2):244-251. doi:10.1017/bhj.2023.25

140 DSA, note 11, Art. 36 (1) and (2).

141 DSA, note 11, Recital 59.

142 DSA, note 11, Recitals 114 and 145, and Art. 14.

143 DSA, note 11, Art. 21.

144 Digital Services Act Observatory, 'The Out-of-court Settlement Mechanism under the DSA: Questions and Doubts (2023) https://dsa-observatory.eu/2023/10/26/the-out-of-court-settlement-mechanism-under-the-dsa-questions-and-doubts/ (accessed 29 May 2024).

Furthermore, in the context of hate speech, this Chapter defends that the DSA should have clarified that online platforms, with a particular emphasis on VLOPs due to its systemic risks, which caused or substantively contributed to criminal hate speech have to comply with corporate remedial responsibilities. These corporate remedial responsibilities are heightened in the case of criminal hate speech amounting to gross human rights violations.

5.5.1.3 Complementary corporate remedial frameworks for hate speech online

At the European level, there is one legal and two policy instruments that complement the corporate remedial framework in the DSA applicable to hate speech on online platforms *i.e.*, respectively, the 2018-revised AVMSD, the Code of Conduct on countering illegal hate speech online, and the Recommendations CM/Rec(2022)16 and CM/Rec(2014)6.

The 2018-revised AVMSD prescribes the State's obligation to regulate inter alia video-sharing platforms with the goals of protecting children and consumers, combating racial and religious hatred, safeguarding media pluralism.[145] In the AVMSD, video-sharing platforms include online platforms disseminating user-generated videos with the purpose to inform, entertain, or educate, and where content organization is decided by the video-sharing platform.[146] Art. 28b of the AVMSD addresses businesses directly and establishes the corporate human rights responsibilities of video-sharing platforms to moderate content.[147] Analysing the remedial responsibilities framework in the AVMSD, Art. 28b(3)(i) clarifies that video-sharing platforms should establish "easy-to-use" complaints mechanisms.[148] Regarding the remedial outcomes, the AVMSD 2010 version had included a specific remedial outcome for audiovisual media services *i.e.*, the right of reply.[149] Nevertheless, the 2018-revised AVMSD did not clarify whether this provision applies to video-sharing platforms.[150]

The Code of conduct on countering illegal hate speech online was agreed in 2016 between the European Commission and internet intermediaries, some of which qualifying as VLOPs as per the DSA.[151] This co-regulatory instrument establishes minimum transparency requirements for content moderation

145 AVMSD, note 12.
146 AVMSD, note 12, Art. 1(1)(b)(aa).
147 AVMSD, note 12, Arts. 28a and 28b.
148 AVMSD, note 12, Art. 28b (i).
149 AVMSD 2010/13/EU, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32010L0013 (accessed 29 May 2024).
150 AVMSD 2010/13/EU, note 150, Recital 103 clarifies that the right to reply can apply online; see also Art. 28.
151 European Commission (2016) The CoC on countering illegal hate speech online; European Commission, 'DSA: Commission designates first set of VLOPs and Search Engines' (2023) https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413 (accessed 29 May 2024).

aiming to counter online hate speech which include clear communication to the users regarding the processes to notify, review, and request removal of hate speech. Notwithstanding, similarly to the DSA, the Code of conduct does not provide a comprehensive framework on corporate remedial responsibilities, processes, or outcomes required from online platforms which have caused or contributed to hate speech.

CM/Rec(2022)16 reiterates the right to an effective remedy,[152] and clarifies that remedial processes should be accessible through civil, administrative, and out-of-court mechanisms.[153] Additionally, CM/Rec(2022)16 explains that some of the most adequate remedial outcomes for online hate speech include: compensation, deletion, blocking, injunctive relief, publication of an acknowledgment that a post constituted hate speech, fines, and loss of licence.[154] Reviewing the CM/Rec(2022)16 corporate remedial standards against the UNGPs, it becomes clear that, though it expands on remedial processes and outcomes, CM/Rec(2022)16 missed an opportunity to distinguish between the State's duty to ensure access to the right to remedy and the corporate remedial responsibilities of online platforms.

CM/Rec(2014)6 elaborates on human rights for internet users and advances that, for criminal acts committed online, the most effective remedies include *inter alia* an inquiry, an explanation by the service provider, the possibility to reply, reinstatement of user-created content, reconnection to the Internet, and compensation.[155] Similarly to the CM/Rec(2022)16, this is an important analysis of the suitability of remedial outcomes for online criminal acts which sheds light on the application of the UNGPs remedial framework on online platforms.

Overall, despite occasional references in the European regulatory framework to the corporate remedial responsibilities of online platforms, these instruments lack a comprehensive approach to the framework on corporate remedial responsibilities, processes, and required outcomes for online platforms that caused or contributed to criminal hate speech.

5.5.2 Proposed standards for a comprehensive framework

This section proposes standards to address the existing loopholes and for a comprehensive framework on the modes of responsibility, remedial processes, and remedial outcomes applicable to online platforms which caused or contributed to criminal hate speech. These standards build on the general framework stemming from the UNGPs on corporate remedial responsibilities.

152 CM/Rec(2022)16, note 17, Para. 20.
153 CM/Rec(2022)16, note 17, Paras. 75 and 90.
154 CM/Rec(2022)16, note 17, Para. 75.
155 CM/Rec(2014)6, note 22, Para. 103.

Regarding the modes of responsibility, the European regulatory framework should clarify, in a consistent manner, that online platforms, with an emphasis on VLOPs as per the DSA, which caused or contributed to adverse human rights impacts are responsible for providing remediation. Hence, this remedial responsibility applies for online platforms which caused or contributed to criminal hate speech. This can be achieved, for example, through the development of an additional chapter in the DSA. The clarification of the modes of responsibility are all the more important in cases where the online platform caused or contributed to criminal hate speech amounting to gross violations of human rights. For cases where the online platform was directly linked to actual or potential but avoidable dissemination of criminal hate speech, they should use their leverage to prevent or mitigate said criminal hate speech.

Vis-à-vis the remedial processes, the European regulatory framework should clarify that remedial processes ought to be legitimate, prompt, and impartial in addressing the adverse human rights impacts, including the dissemination of criminal hate speech on online platforms. Though the DSA standardizes operational-grievance mechanisms such as the internal appeals and transparency standards, the European legislators should ensure that remedial processes apply human rights standards and not terms and conditions privately decided by online platforms and often in misalignment with human rights.

Concerning the remedial outcomes, the European regulatory framework fails to establish a clear and comprehensive approach to corporate remedial outcomes required of online platforms which caused or contributed to adverse human rights impacts, including for cases of criminal hate speech and criminal hate speech amounting to gross human rights violations. The following subsections explore the suitability of remedial outcomes[156] by building on the framework of remedial outcomes for criminal acts online. The theoretical frameworks for remedial outcomes are: restitution and satisfaction as amplification of survivors' speech (Section 5.5.2.1); compensation and rehabilitation beyond the area of services (Section 5.5.2.2); and, guarantees of non-repetition as business models' change (Section 5.5.2.3). These remedial outcomes could be imposed by the European Commission as interim non-judicial measures applicable to online platforms which caused or contributed to criminal hate speech.[157]

For the overall operationalization of these standards, this Chapter recommends that the European Commission issues a detailed guidance on Art. 21 of the DSA in alignment with the UNGPs corporate human rights remedial responsibilities framework. Such guidance should explicitly clarify the modes of responsibility, remedial processes, and remedial outcomes suitable to effectively and promptly remediate people harmed by criminal hate speech dissemi-

156 Section 4.3.
157 DSA (note 11), Art. 70.

nated by online platforms. These standards are all the more urgent to clarify for VLOPs as per the DSA, and for cases of criminal hate speech amounting to gross violations of human rights.

5.5.2.1 Restitution and satisfaction as amplification of survivors' speech

Online platforms which caused or contributed to criminal hate speech must provide for restitution as a means to restore, to the extent possible, the exercise of adverse human rights impacts. In compliance with the standards on satisfaction, businesses must recognize the acts that violated international law and restore the survivors' dignity.[158]

Though there is a vast array of harms resulting from human rights violations in these cases,[159] this section proposes a remedy for the specific harm of constrained online participation. To clarify, some of the most commonly reported harms resulting from online hate speech (and even more so from criminal hate speech) are disempowerment, silencing, and ultimately disengagement from online platforms of targeted communities.[160]

A remedy to the constrained participation of communities targeted by hate speech is the speaking back capabilities framework advanced by Gelber.[161] In this framework, Gelber defends that policy and legal approaches should support people targeted by hate speech who wish to respond to it.[162] This direct engagement in the response process is conceptualized as the empowering act which enables communities targeted by hate speech to overcome the oppression and harm of constrained participation.[163] Gelber explains that this framework can result in policies of affirmative speech in which actors that enabled and hosted hate speech should likewise facilitate the response and counter narratives.[164]

This Chapter expands on Gelber's speaking back framework by applying it to the context of online platforms. Importantly, it is widely discussed how the harm caused by hate speech is aggravated by online platforms when their

158 Stoica (note 118), 148.
159 Section 3.1.
160 Katharine Gelber (2002). Speaking Back. The free speech versus hate speech debate. John Benjamins Publishing Company, 117, 118.
161 Gelber (note 160).
162 This research builds on decolonial and feminist sociology and psychology theories as well as empirical studies showing that the direct engagement and leadership of people targeted by hate speech in deciding the response to the harm caused empowers and contributes to a faster overcoming of the oppression perpetuated by hate speech.
163 Gelber (note 160), 119.
164 Gelber (note 160), 124.

algorithms demote counter narratives.[165] In this context, this Chapter suggests that online platforms which caused or contributed to criminal hate speech should, as an effective restitution remedy, introduce affirmation speech policies in their content ranking, moderation, and recommendation algorithms.

As a result, for a given period, online platforms should amplify survivors' speech through content ranking algorithms. Similarly, online platforms should deploy content moderation algorithms that will specifically detect and apply a higher scrutiny to hate speech posts targeting marginalized communities with the goal of avoiding double victimization. Finally, to ensure reconnection of marginalized people as groups, online platforms should adopt affirmative speech policies through their link recommendation algorithms by purposefully, for a given period, connecting people marginalized and targeted by such criminal hate speech.

5.5.2.2 Compensation and rehabilitation beyond area of services

Online platforms which caused or contributed to criminal hate speech should remediate psychological, physical, and material harms through rehabilitation and compensation. This overarching remedial responsibility clarified that online platforms are responsible to remediate survivors beyond their area of services.

For cases of criminal hate speech amounting to gross human rights violations, online platforms are explicitly required to ensure access and, importantly, pay for rehabilitation and compensation of medical and psychological services. Moreover, in these cases, the conceptualization of victims expressly includes not only the directly affected persons but also others closely related. Finally, the European Commission may impose asset freezing on online platforms which caused or contributed to criminal hate speech amounting to gross human rights violations.[166]

Applying these remedies to the example of Meta's significant contribution to the genocide of the Rohingya in Myanmar, it becomes clear that Meta should allocate funds and has the corporate remedial responsibility to compensate and rehabilitate beyond its area of service. This responsibility should address material harms including lost opportunities such as limited access to employment or education.

---

165 E.g., Oliver L. Haimson et al. "Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-35; Daniel Delmonaco et al. "" What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and" Prove" Shadow banning." Proceedings of the ACM on Human-Computer Interaction 8.CSCW1 (2024): 1-39.

166 European Union (note 99).

5.5.2.3 Guarantees of non-repetition as business models' change

Many online platforms have adopted business models, as well as designed and deployed content moderation, ranking, and recommendation algorithms that maximize profit and user engagement often at the expense of human rights.[167] All online platforms have the corporate human rights responsibility to identify, prevent, mitigate, and remediate adverse human rights impacts.[168] Online platforms which caused or contributed to adverse human rights impacts criminal hate speech have the heightened responsibility to remediate, including by adopting guarantees of non-repetition.

This Chapter proposes the operationalization of guarantees of non-repetition premised on a change of business models and grounded in two main elements: (1) enforcing content moderation, ranking, and recommendation algorithms based on human rights standards; (2) enforcing an alignment of the terms of service with the international human rights standards on the conceptualization of criminal hate speech and with the corporate human rights responsibilities framework in the UNGPs.

First, online platforms should ensure that their content moderation algorithms remove criminal hate speech. Notably, as per Art. 5(1)(a) of the AI Act, online platforms are prohibited from deploying algorithms that are likely to lead to violence, as is the case of criminal hate speech. A key provision in verifying compliance with these responsibilities is Art. 40 of the DSA, which enables researchers to access data from VLOPs to investigate the impact of algorithms on systemic risks, including hate speech. In this context, this Chapter suggests that, when assessing compliance with Art. 5 of the AI Act (in non-judicial or judicial actions), the judicial burden of proof should be inverted to require online platforms to prove that they did not cause nor contributed to criminal hate speech.[169] Though this inversion of the burden of proof is not clarified within the CSDDD, this Chapter proposes that this is a key means for online platforms to comply with their duty of care.[170]

Regarding ranking and recommendation algorithms,[171] this Chapter builds on two contextual variables utilized by the ECtHR to assess the severity of hate speech[172] to suggests a tighter framework for monitoring criminal hate speech i.e., the political and social background, as well as the speaker's status

167 Introduction and Section 2.2.
168 UNGPs (note 7), Principle 15.
169 In this context, it is important to adequately identify and mitigate potential complicated implications for national criminal law procedures.
170 Caio CV Machado and Thaís Helena Aguiar. "Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models." Business and Human Rights Journal 8.2 (2023): 244-251.
171 Amnesty International (note 5), 8, highlights that "content moderation alone is inherently inadequate as a solution to algorithmically-amplified harms."
172 Section 2.2.

or role in society. This Chapter suggests that, as a minimum legal standard especially during times of conflict or elections, online platforms should proactively monitor users and posts with high levels of engagement above a certain threshold of risk. The notion of engagement level expands on Gelber's authority framework, whereby measuring authority of a certain speech-act is relevant to analyse the capability of harming.[173] In this context, the engagement level corresponds to the notion of authority and could track two parameters i.e., the number of followers of a given user or the number of reactions (e.g., reposting, comments) to a given post.

Secondly, online platforms should reflect the corporate human rights responsibilities framework in their terms of service as instructed in the UNGPs and in the CSDDD, including by adopting a conceptualization of criminal hate speech aligning with international human rights standards.[174] Furthermore, online platforms should transparently inform users about the proposed content moderation, ranking, and recommendation standards, as well as the tighter contextual application during conflicts or elections. Finally, as a minimum legal standard, after detection of criminal hate speech, online platforms should be required to archive such content for future criminal investigations.[175]

5.6 Conclusion

This Chapter addresses the key challenge of the lack of legal clarity about the corporate remedial responsibilities of online platforms that caused or contributed to criminal hate speech. The research question is two-fold: To ensure the right to an effective remedy, how can European legislators better align the legal framework on the corporate remedial responsibilities of online platforms which caused or contributed to criminal hate speech in order to better align it with the general framework on corporate remedial responsibilities? Additionally, are there heightened remediate responsibilities for very large online platforms or for cases of criminal hate speech amounting to gross violations of human rights?

By building upon the European conceptualization of criminal hate speech, the European standards on the right to an effective remedy, and the general framework of corporate human rights responsibilities, this Chapter proposes

173 Gelber (note 30), 401.

174 In misalignment with human rights, Facebook's terms of service allows hate speech towards criminals. This was one of the criteria permitting hate speech towards two members of the Rohingya community who were initially wrongly accused of having raped. E.g., Nave, note 42.

175 This research acknowledges the growing records of infiltration of extremists in law enforcement bodies. Daniel Koehler, 'From superiority to supremacy: Exploring the vulnerability of military and police special forces to extreme right radicalization' Studies in Conflict & Terrorism (2022): 1-24.

three legal avenues for the European legislators to clarify the framework on corporate remedial responsibilities.

First, it is important to clarify that the individual right to an effective remedy results in, not only a State obligation to ensure the exercise of said right, but also in direct corporate remedial responsibilities. Second, the corporate remedial responsibilities framework must address: remedial responsibilities modes; remedial processes; and, remedial outcomes. Third, the corporate remedial outcomes must be tailored to address the specific harms caused by criminal hate speech online through content moderation, ranking, and recommendation algorithms.

Delving deeper onto the most effective remedial outcomes for criminal hate speech, this Chapter suggests the amplification of survivors' speech as means to restore the harm of limited participation. For the remaining harms, online platforms should compensate and rehabilitate beyond their area of services. Finally, this Chapter suggests that the only way in which online platforms can remediate through guarantees of non-repetition of harm is by ensuring that their business model prioritizes human rights over profit.

The standards proposed in this Chapter on corporate remedial responsibilities apply to online platforms, with increased corporate human rights responsibilities for VLOPs and platforms which caused or contributed to elements of criminal hate speech amounting to gross violations of human rights. These suggested legal avenues apply first and foremost to the European context given the existing regulatory framework clarifying the conceptualization of criminal hate speech, particularly since the adoption of CM/Rec(2022)16.

Importantly, the interventions to counter criminal hate speech on online platforms should not be solely legalistic nor should they just rely on remedy after the adverse impact on human rights has occurred. There should be structural changes to addressing power imbalances and systems of privilege, namely through education, representation, and through the regulation of the private sector prioritizing profit over human rights.