Universiteit
Leiden
The Netherlands

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

Nave, E.V.R.

Citation
Nave, E. V. R. (2025, July 3). *Countering online hate speech: how to adequately protect fundamental rights?*. *Meijers-reeks*. Retrieved from https://hdl.handle.net/1887/4252655

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4252655

Note: To cite this publication please use the final published version (if applicable).

3 Human rights responsibilities of online platforms to prevent criminal hate speech

How do european corporate preventive human rights responsibilities impact terms of service?[12]

ABSTRACT

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

1 This Chapter was originally published in the Computer Law and Security Review 51: 105884, in co-authorship with Lottie Lane, Assistant Professor of Public International Law, University of Groningen.

2 This Chapter was updated after publication and hence the content deviates from what was previously published. More specifically, 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.

communities by aligning their ToS with the conceptualisation of criminal hate speech in European human rights standards.

3.1 INTRODUCTION

Around two thirds of the world's population are active Internet users.[3] While the Internet enables individuals to access information and exercise their freedom of expression, it also enables the proliferation of online hate speech. In this Chapter, we assess whether online platforms[4] could be required, as part of emerging European human rights due diligence (HRDD) responsibilities, to align their Terms of Service (ToS)[5] with the conceptualisation of criminal hate speech[6] in European human rights standards.

'Online hate speech' broadly refers to discriminatory expressions shared through the Internet targeting historically marginalised[7] people based on their inherent characteristics. Recommendation CM/Rec(2022)16 adopted by the Council of Europe (CoE) Committee of Ministers (CM) in May 2022[8] explains that hateful expressions represent a violation of human rights. When unaddressed, these can hinder peace and development by denying the values of pluralism, tolerance and broadmindedness essential in a democratic society.

The rise of online hate speech results from specific features of the Internet. First, unlike in traditional media, most content published on the Internet can be quickly shared with little to no monitoring, made available to large audiences, published under anonymity, and easily manipulated in ways that intensify hate (e.g. hate profiles, memes and deep fakes). Second, online content is hosted by businesses primarily driven by profit goals, often at the expense of human rights. The potentially negative impact of AI-driven content moderation by online platforms is under increasing scrutiny. For example, Meta Platforms, Inc. (formerly named Facebook, Inc.) faces legal action for alleged

3 Number of internet and social media users worldwide as of January 2023 (2023), available at <https://www.statista.com/statistics/617136/digital-population-worldwide/> accessed 19 November 2024.

4 In alignment with the terminology in the Digital Services Act, this Chapter uses 'online platforms' to refer to social media platforms. Where we discuss the broader framework of corporate human rights due diligence applicable to artificial intelligence (AI) businesses more generally, we use 'AI businesses'; we consider online platforms to be a sub-category of AI businesses. We use 'businesses' and 'companies' interchangeably.

5 This Chapter uses ToS, 'Community Guidelines' and 'Terms and Conditions' interchangeably to refer to policy communications between the companies and their users laying down the rules of engagement with the AI service.

6 This Chapter follows the European human rights standards stemming from the jurisprudence of the ECtHR by using interchangeably 'criminal hate speech' and 'the most serious forms of hate speech'.

7 This Chapter uses 'oppression' and 'marginalisation' interchangeably.

8 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).

negligence in facilitating the genocide of Rohingya Muslims in Myanmar after its algorithm failed to remove hateful posts and amplified hate speech.[9] Similarly, whistle-blower Frances Haugen alerted that Facebook neglected reports of accounts and hate speech content towards Muslims in India, potentially leading to offline violence.[10] There are reportedly other situations of human rights abuses by different platforms.[11]

Legal scholars have alerted to the growing impact of social media platforms on the application of regulatory frameworks for freedom of expression and democratic processes, and to the subsequent need to expand the legal scholarship focusing on the regulation of online platforms.[12] In this context, it is relevant to consider that most of these online platforms are based in the USA and thus typically bound by the USA framework on freedom of expression, corporate human rights due diligence and intermediary liability. Conversely, to the extent that these online platforms operate in European Union (EU) territory, they must also abide by the regional human rights frameworks in Europe, which differ significantly from those applicable in the USA. The reconciliation of different regional standards has been challenging, not only for online platforms but also for judicial bodies in enforcing their decisions.[13]

---

9 Al Jazeera, 'Rohingya sue Facebook for $150bn for fuelling Myanmar hate speech' (7 December 2021), available at <https://www.aljazeera.com/news/2021/12/7/rohingya-sue-facebook-for-150bn-for-fuelling-myanmar-hate-speech> accessed 6 April 2023.

10 Al Jazeera, 'Facebook failing to check hate speech, fake news in India: Report' (25 October 2021), available at <https://www.aljazeera.com/news/2021/10/25/facebook-india-hate-speech-misinformation-muslims-social-media> accessed 6 April 2023.

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

12 E.g. Kate Klonick, 'The new governors: The people, rules, and processes governing online speech'(2017) *Harv. L. Rev.*, *131*, 1598; Tarlach McGonagle, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation' (2020) Oxford Handbooks in Law; Giancarlo Frosio (Ed.) (2020) *Oxford handbook of online intermediary liability*, Oxford Handbooks. (pp. 467–485), 10; Tarlach McGonagle, 'The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing', 232, in Rikke Frank Jørgensen (eds), '*Human Rights in the Age of Platforms*' (2019) Cambridge, MA: The MIT Press, available at <https://doi.org/10.7551/mitpress/11304.001.0001> accessed 19 November 2024; Judit Bayer, Bernd Holznagel, Päivi Korpisaari (ex. Tiilikka), Lorna Woods (2021) Baden-Baden: Nomos Verlagsgesellschaft mbH & Co. KG., Volume 1, 30, available at <https://doi.org/10.5771/9783748929789> accessed 19 November 2024; Martin Moore and Tambini Damian (eds), 'Regulating Big Tech: Policy Responses to Digital Dominance' (2021), available at <https://doi.org/10.1093/oso/9780197616093.001.0001> accessed 19 November 2024.

13 E.g. Ligue contre le racisme et l'antisémitisme et Union des étudiants juifs de France c. Yahoo! Inc. et Société Yahoo! France (LICRA v. Yahoo!). and Yahoo Inc. v LICRA; European Court of Justice, Opinion of Advocate General Szpunar delivered on 8 June 2023 (1) Case C-376/22 clarifies that Union law prescribes the possibility for Member States to restrict

Legislators and policymakers at the international, regional and national level have made many efforts to prevent and address the negative impact of business on human rights, including through HRDD and through liability regimes. The HRDD regime includes the seminal United Nations Guiding Principles on Business and Human Rights (UNGPs), which are arguably the most authoritative international expression of the corporate responsibility to respect human rights through HRDD.[14] At the European Union (EU) level, a Directive on corporate sustainability due diligence (CSDDD) was just recently adopted.[15] Businesses – including online platforms – falling under the scope of the CSDDD should identify, prevent, mitigate and bring an end to negative impacts on human rights. Furthermore, the EU adopted the Artificial Intelligence Act (AI Act) emphasising the need for protection of human rights in the digital environment.[16]

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

---

the freedom to provide information society services to 'fight against any incitement to hatred on grounds of race, sex, religion or nationality, and violations of human dignity concerning individual persons'.

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

15 European Union (2024) Directive (EU) 2024/1760 of the European Parliament and of the Council of 13 June 2024 on corporate sustainability due diligence and amending Directive (EU) 2019/1937 and Regulation (EU) 2023/2859.

16 Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), Explanatory Memorandum, 1.1.

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

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

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

20 European Commission (2016) The CoC on countering illegal hate speech online.

Despite these advancements, the HRDD framework applicable to online hate speech has focused mostly on explaining the responsibilities of companies throughout their operations. Guidance regarding HRDD requirements for the regulation of hate speech in the ToS is missing. A key aspect remains unaddressed: how online businesses should define hate speech and how this should be communicated to their users. More specifically, is there a legal standard emanating from the European HRDD framework prescribing the responsibility for online platforms[21] to align their ToS, 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 answer this research question, Section 3.2. employs doctrinal research to clarify the elements of the most serious cases of hate speech regulated by criminal law. The legal framework of criminal hate speech presents a common European understanding under which specific HRDD can be required of online platforms. As explained in CM/Rec(2022)16, the most serious cases of hate speech represent a criminally actionable violation of rights under Article 17 of the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR).[22] The emphasis on criminal hate speech is particularly important since the European Commission (EC) proposed an extension of the list of EU crimes to include hate speech.[23]

In Sections 3.3. and 3.4, this Chapter deals with the HRDD regime.[24] Section 3.3. explores the HRDD framework applicable to AI businesses.[25] The key instruments analysed are the UNGPs, Organization for Economic Cooperation

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

22 Recommendation CM/Rec(2022)16 (n 8) paragraph 7 and Explanatory Memorandum, paragraph 27. See Françoise Tulkens, 'When to say is to do: Freedom of expression and hate speech in the case-law of the European Court of Human Rights' in Josep Casadevall, Egbert Myjer, Michael O'Boyle, and Anna Austin (eds), Freedom of Expression: Essays in honour of Nicolas Bratza (Wolf Legal Publishers, 2012) 284.

23 European Commission, 'The Commission proposes to extend the list of "EU crimes" to hate speech and hate crime' (9 December 2021), available at <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_6561> accessed 6 April 2023.

24 As a regulatory approach distinct from that of HRDD – as seen in the separate chapters on each regime in the DSA –, the EU liability regime for internet service providers (ISP) falls outside the remit of this research. These regimes are nevertheless related in that liability may follow from non-compliance with HRDD responsibilities. For discussion of ISP liability regimes and recent case law, see e.g. Andrea Bertolini et al., 'Liability of Online Platform: Study for the European Parliament' (2021) European Parliamentary Research Service PE 656.318; Berrak Genç-Gelgeç, 'Regulating Digital Platforms: Will the DSA Correct Its Predecessor's Deficiencies?' (2022) 18 Croatian Yearbook of European Law and Policy 25; United States Supreme Court, Twitter v. Taamneh *598 US (2023).*

25 AI businesses are companies that provide services based on artificial intelligence methods and include inter alia online platforms and thus are a relevant framework for the analysis in this Chapter.

and Development (OECD) initiatives, and the EU CSDDD and AI Act. International instruments are included because they provide a substantive understanding of corporate responsibility for human rights that has influenced the development of the CSDDD and can provide inspiration regarding how European instruments should be interpreted. Doctrinal research is used to identify and address legal loopholes from a human rights perspective.

Section 3.4. delves deeper into preventive HRDD responsibilities in moderation of illegal content, such as criminal hate speech. The legal instruments examined are the DSA, the AVMSD, the CoC[26] and CM/Rec(2022)16. Emphasis is placed on HRDD responsibilities in the drafting or updating of the ToS as a means for online platforms to respond to the systemic risk of online hate speech. It is suggested that to improve legal coherence in countering online hate speech in the European context, online platforms should follow CM/Rec(2022)16's conceptualisation of criminal hate speech in their ToS.

Section 3.5. presents an empirical qualitative analysis of three case studies: Facebook,[27] Twitter,[28] and YouTube. We assess the compliance of the platforms' ToS with the European Court of Human Rights (ECtHR) jurisprudence on criminal hate speech, and with the conceptualisation of criminal hate speech in CM/Rec(2022)16. The platforms were selected because they: (1) fall under the scope of CSDDD; (2) are signatories to the CoC; and (3) qualify as very large online platforms (VLOPs) as defined in the DSA.[29]

In summary, this Chapter applies the European HRDD framework of online platforms to the conceptualisation of criminal hate speech in ToS. The main finding is the proposal of a minimum HRDD legal standard that online platforms operating in Europe must align their ToS with the European human rights conceptualisation of the most serious cases of hate speech. The EC should issue a sector-specific guidance suggesting the adoption of such legal standard.

3.2 Online hate speech is always illegal, sometimes criminalised

This section lays the theoretical framework regarding the conceptualisation of hate speech grounding the subsequent discussions of corporate HRDD

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

27 Owned by Meta Platforms Meta.

28 Now X. For the purpose of coherence, reference is made to the company name at the time of the study.

29 *I.e.*, they have 45 million or more average monthly active recipients of their service in the Union: DSA, Recital 76.

responsibilities.[30] We clarify the key elements in the original conceptualisation of hate speech in critical race theory (Section 3.2.1) and explain key conceptual elements in the European regulatory framework countering hate speech (Section 3.2.2).

3.2.1 Original conceptualisation

The term 'hate speech' became prominent in the work of critical race scholars in reference to 'racist hate speech'.[31] Scholars like Mari J Matsuda emphasised that racist hate speech is used to perpetuate the marginalisation of historically oppressed groups and thus should not be protected under the right to freedom of expression.[32] Matsuda conceptualises three elements in racist hate speech:

> '1) the message is of racial inferiority and all members of the target group are considered alike and inferior; 2) the message is directed against a historically oppressed group and reinforces a historically vertical relationship; 3) the message is persecutory, hateful and degrading'.[33]

Hate speech can cause different harms, including physical, psychological and socio-economic harm.[34] For example, people targeted by hate speech may develop low self-esteem, post-traumatic stress disorder, psychosis or depression.[35] Critical legal scholars have also stressed the effect of cumulative exposure to hate speech, as people targeted by hate speech on a continuous basis may experience particularly severe psychological harm,[36] and may have more difficulties in succeeding at their studies or at their jobs, as they may withdraw from society to avoid such hateful messages.[37]

A key analytical framework presented by critical race scholars to understand the harms caused by hate speech is the intersectionality of systems of marginalisation. Kimberlé Crenshaw underlines the importance of examining the intersection between different types of discrimination by highlighting how

30 This section summarises the main argument in Eva Nave, 'Hate speech, historical oppression and European human rights (2023) Buffalo Human Rights Law Review.

31 Critical race scholars contest neutral viewpoints in research and highlight the impact of institutional inequalities deriving from moments when colonial and discriminatory doctrines were openly defended.

32 Mari J Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 Michigan Law Review 2320, 2335.

33 Ibid 36.

34 See Richard Delgado and Jean Stefancic, *Understanding Words That Wound* (Routledge 2004) 12–19, available at <https://doi.org/10.4324/9780429503351> accessed 19 November 2024.

35 Ibid 14.

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

37 Delgado and Stefancic (n 34), 111–121.

the politics of race and gender marginalise racialised women,[38] and exposed the shortfalls of legal and political approaches that isolate systems of oppression. Though the intersectionality theory was initially developed considering systems of marginalisation based on racial and gender markers, Crenshaw explained that the theory applies to the intersection of any system of marginalisation such as class, sexual orientation and age.[39]

3.2.2 Key conceptual elements in European regulation

There is currently no legally binding definition of hate speech in international or European human rights law. However, both the CoE and the EU have developed legal strategies to counter hate speech by clarifying key elements in the conceptualisation of hate speech or explaining the procedural responsibilities of stakeholders involved in the moderation of speech (e.g. media, Internet intermediaries,[40] law enforcement, governments). Though this section identifies the main instruments regardless of the type of strategy, we focus on the instruments that expand on the key conceptual elements of hate speech.

There is an overall alignment of key human rights values between the two European systems. To ensure European legal consistency, Article 52(3) of the Charter of Fundamental Rights of the EU (CFREU)[41] requires the same meaning and scope to be given to CFREU provisions as to corresponding rights in the ECHR. Furthermore, in Article 6(2) of the Treaty of the European Union (TEU) the EU commits to acceding to the ECHR,[42] which will enable individuals to submit to the ECtHR complaints of violations of ECHR by the EU.[43] Thus, both the CoE and the EU constitute reference systems to summarise the main elements of the European regulation of hate speech.

38 Kimberlé Crenshaw, 'Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color' (1990) Stanford Law Review 1241, 1243.

39 Mari J Matsuda, Charles R. Lawrence III, Richard Delgado and Kimberle Williams Crenshaw, *Words That Wound Critical Race Theory, Assaultive Speech, And The First Amendment* (Routledge, 1993) 114.

40 'Internet intermediaries' includes online platforms.

41 CFREU, Article 52(3).

42 The accession negotiations of the EU to the ECHR resumed in 2020.

43 Each of the EU member states is already party to the ECHR. However, without the EU's accession to the ECHR, individuals cannot lodge complaints against EU institutions. The accession will mean that the EU will be subjected to the oversight of the ECtHR in the application of the ECHR. Further information at 'European Union Accession to the European Convention on Human Rights – Questions and Answers', available at <https://www.coe.int/en/web/portal/eu-accession-echr-questions-and-answers> accessed 6 April 2023.

At the EU level, there are strategies to counter hate speech in primary, secondary and 'soft' law sources.[44] While some strategies focus on substantive regulation (i.e. the conceptualisation of hate speech), most focus on procedural regulation (e.g. the liability of internet intermediaries, HRDD). Though the next paragraphs summarise the main strategies, Internet intermediaries' HRDD responsibilities are addressed more thoroughly in Sections 3.3. and 3.4. Importantly, this Chapter does not focus on intermediary liability, but rather on human rights due diligence responsibilities.

Following the abovementioned alignment of primary sources of EU law with the ECHR,[45] content in the provisions in the CFREU addressing hate speech should be interpreted in the same way as the ECtHR interpretation for the equivalent provisions in the ECHR. The most relevant secondary legal sources are the Council Framework Decision on combating *certain* forms and expressions of racism and xenophobia by means of criminal law (Framework Decision),[46] the AVMSD,[47] the DSA.[48] Finally, the main supplementary legal source at the EU level is the CoC.

Despite the variety of EU regulatory strategies applicable to hate speech, there is no coherent and all-encompassing framework. On the one hand, the CoC and the DSA refer to the conceptualisation of hate speech as presented in the Framework Decision which focuses only on *certain* forms of racist and xenophobic hate speech (by reference to race, colour, religion, descent or national or ethnic origin), thus excluding other types of hate speech such as misogyny and queerphobia. This is all the more worrisome as data presented in the latest monitoring round of the CoC indicate that hatred on accounts of sexual orientation is the most commonly reported ground for hate speech.[49] On the other hand, the AVMSD applies a different legal rationale, referring

44 Primary legal sources of EU law are the treaties establishing the EU and general legal principles. Secondary sources of EU law comprise legislative delegated and implementing acts. Further information on EU legal sources available at <https://www.europarl.europa.eu/factsheets/en/sheet/6/sources-and-scope-of-european-union-law> accessed 19 November 2024. 'Soft law' refers to non-legally binding sources that may aid the interpretation of hard, binding law and which may have an impact on businesses' behaviour in practice. For further information in the field of business and human rights, see Sarah Joseph and Joanna Kyriakakis, 'From soft law to hard law in business and human rights and the challenge of corporate power' (2023) 36(2) LJIL 335.

45 TEU, Article 6(2).

46 [emphasis added]. Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law 2008, OJ L 328.

47 AVMSD (n 18).

48 DSA (n 17). There are also three legislative instruments: the Regulation of the EP and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act, AI Act); the Directive on adapting non contractual civil liability rules to artificial intelligence; and, the EC Proposal for a Directive of the EP and of the Council on combating violence against women and domestic violence.

49 CoC 7th monitoring round report (2022), 4.

to the broader list of grounds of prohibited discrimination in Article 21 CFR, 'such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation'.

The CoE level also developed various regulatory approaches to counter online hate speech, including legally binding and non-binding initiatives. The most relevant treaties are the ECHR, the 2003 Additional Protocol to the Convention on Cybercrime,[50] the 2011 Convention on preventing and combating violence against women and domestic violence (the Istanbul Convention),[51] and the 1994 Framework Convention for the Protection of National Minorities.[52] The most relevant non-binding initiatives include Recommendations[53] by the CM and General Policy Recommendations of the European Commission against Racism and Intolerance (ECRI).[54] The work by the CM and ECRI is essential to understand new phenomena and is frequently cited in the ECtHR's jurisprudence.[55]

The ECtHR has developed an extensive body of jurisprudence interpreting the ECHR and referring to various CoE instruments regulating hate speech, the most relevant of which is the ECHR. The main provisions cited by the ECtHR have been the prohibition of abuse of rights, the provision containing

50 Council of Europe, Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems, ETS 189.

51 Council of Europe Convention on preventing and combating violence against women and domestic violence, ETS No. 210.

52 CoE, Framework Convention for the Protection of National Minorities and Explanatory Report, H (95) 10.

53 The most relevant CM recommendations include: (1997)20 on hate speech; (1997)21 on the media and the promotion of a culture of tolerance; (2010)5 on measures to combat discrimination on grounds of sexual orientation or gender identity; (2011)7 on a new notion of media; (2016)3 on human rights and business; (2018)2 on the roles and responsibilities of internet intermediaries; (2020)1 on the impact of algorithmic systems on human rights; and, (2022)16 on a wide-ranging strategy to combat hate speech in light of current challenges. Aside from the recommendations, the CM also adopted a 'Declaration on the manipulative capabilities of algorithm processes' (2019) and 'Guidelines on upholding equality and protecting against discrimination and hate during times of crisis' (2021).

54 The most relevant GPR of ECRI include: GPR No. 6 on combating the dissemination of racist, xenophobic and antisemitic material via the Internet; GPR No. 7 on national legislation to combat racism and racial discrimination; GPR No. 11 on combating racism and racial discrimination in policing; and, GPR No. 15 on combating hate speech.

55 Tarlach McGonagle, 'The Council of Europe against Online Hate Speech: Conundrums and Challenges', (MCM; No. 2013(005)), Council of Europe Conference of Ministers responsible for Media and Information Society "Freedom of Expression and Democracy in the Digital Age: Opportunities, Rights, Responsibilities" 40, 44 (2013), available at <http://www.coe.int/t/dghl/standardsetting/media/Belgrade2013/McGonagle%20- %20The%20Council%20of%20Europe%20against%20online%20hate%20speech.pdf> accessed 6 April 2023; Keynote speech of Nils Muiz?nieks, 'Freedom of Expression and Democracy in the Digital Age: Opportunities, Rights, Responsibilities' (Council of Europe, 2013), available at <https://www.coe.int/en/web/freedom-expression> accessed 6 April 2023, 27.

the legal requirements for restrictions of freedom of expression,[56] respect for private life, the prohibition of discrimination and the right to an effective remedy.[57] The key elements in the regulation of hate speech in the ECtHR's jurisprudence are found on applications by perpetrators of hate speech. The ECtHR classifies hate speech in two categories: (1) no clear abuse of rights but prohibited under civil or administrative law, as long as the prohibition aligns with Article 10(2) (Section 3.2.2.1); and (2) clear abuse of rights under Article 17 and thus criminally actionable (Section 3.2.2.2). The following subsections expand on these two categories of hate speech by making reference to the CM/Rec(2022)16. CM/Rec(2022)16 is the most significant instrument of the CM building on the ECtHR's jurisprudence and presenting a comprehensive framework to counter online hate speech.

3.2.2.1 Hate speech is always illegal

The first category of hate speech is when, though not criminally actionable, the speech is still prohibited under civil or administrative law. This prohibition of speech needs to align with the criteria emanating from Article 10(2) ECHR, i.e. it should be: i) prescribed by law; ii) in pursuit of one or more specified legitimate interests (national security, territorial integrity or public safety, prevention of disorder or crime, for the protection of health or morals, reputation or rights of others, prevention of the disclosure of information received in confidence, or for maintaining the authority and impartiality of the judiciary); and, iii) necessary in a democratic society.

According to the ECtHR, the necessity test entails an analysis of the following contextual factors:[58] political and social background;[59] intent of the speaker;[60] speaker's status or role in society;[61] content of the expression;[62] extent

56 Françoise Tulkens, 'The Hate Factor in Political Speech: Where Do Responsibilities Lie?', *Report of the Council of Europe Conference, Warsaw 18-19 September 2013* (2013).

57 Articles 17, 10(2), 8, 14 and 13, respectively.

58 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' (2002) 24 Cardozo Law Review 1523, 1565.

59 E.g. *Leroy v France* App No 36109/03 (ECtHR, 02/10/2008); *Ceylan v Turkey* App No 23556/94 (ECtHR, 08/07/1999); *Beizaras and Levickas v Lithuania* App No 41288/15 (ECtHR, 14/01/2020).

60 E.g. *Jersild v Denmark* App No 15890/89 (ECtHR, 09/09/1994).

61 E.g. *Incal v Turkey* App No 22678/93 (ECtHR, 09/06/1998) where the ECtHR ruled that politicians enjoy a protected status but concomitantly have heightened responsibilities to avoid disseminating comments in their public speeches which are likely to foster intolerance. In *Feret v Belgium* App No 15615/07 (ECtHR, 16/07/2009) the ECtHR ruled that politicians have the duty to refrain from using or advocating racial discrimination.

62 E.g. *Goucha v Portugal* App No 70434/12 (ECtHR, 22/03/2016); *Feldek v Slovakia* App No 29032/95 (ECtHR, 12/07/2001); *Ottan v France* App No 41841/12 (ECtHR, 19/04/2018).

of the expression;[63] and the nature of the audience.[64] In this examination of the context, drawing on the insights of critical race and intersectionality theory, it is important to explicitly consider socio-historical marginalisation and the intersectionality of systems of marginalisation affecting people targeted by hate speech.

3.2.2.2 The most serious forms of hate speech are criminally actionable

The second category of hate speech is criminal hate speech.[65] Following the ECtHR's jurisprudence, hate speech is a criminal act when there is a clear abuse of rights under Article 17 ECHR, i.e. when the hateful speech violates or limits (to a greater extent than allowed by the ECHR) any right in the ECHR.[66] These cases of hate speech are considered by the ECtHR to be the most serious forms of hate speech.

Though the ECtHR assesses each application on a case-by-case basis, its jurisprudence on Article 17 reveals the minimum European human rights threshold for hate speech to be considered criminal. This was distilled in Paragraph 11 of the CM/Rec/(2022)16.[67]

CM/Rec(2022)16 seems to suggest an open-ended approach to the conceptualization of impermissible grounds[68] for hate speech by using "such as" when introducing Paragraph 11 before referring to 'racist, xenophobic, sexist and LGBTI-phobic' hate speech.[69] As the ECtHR may in the future be called to rule on serious forms of hate speech targeting people on the basis of their queerness (importantly more broadly conceived than LGBTI[70]), ableism, or non-neurotypical characteristics which, if amounting to the acts explained in Paragraph 11 of CM/Rec/(2022)16, should still be considered an abuse of rights under Article 17 and hence criminal hate speech.

Currently, although no guidance exists at the EU level clarifying the main elements of criminal hate speech, in December 2021, the EC proposed to extend

63 E.g. *Gündüz v Turkey* App No 35071/97 (ECtHR, 04/12/2003) where the ECtHR noted that live TV is not easy to reformulate or retract.

64 E.g. *Vejdeland and others v Sweden* App No 1813/07 (ECtHR, 09/02/2012; *Vereinigung Bildender Künstler v Austria* App No 68354/01 (ECtHR, 25/01/2007).

65 This Chapter uses 'criminal hate speech' and 'the most serious forms of hate speech' interchangeably.

66 Tulkens (n 56) 5.

67 CM/Rec/(2022)16, Paragraph 11.

68 This Chapter uses 'impermissible grounds for hate speech', 'protected categories' and 'protected characteristics' interchangeably.

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

70 The New York Times (2022), Using the Word 'Queer" Instead of 'Gay', available at <https://www.nytimes.com/2022/11/13/opinion/letters/lgbt-gay-queer.html> accessed 22 November 2024; Heather Love, "Queer." *Transgender studies quarterly* 1.1-2 (2014): 172-176.

the list of EU crimes to hate speech.[71] Studying the CoE developments, such as the CM/Rec/(2022)16, to inform the EU regulatory initiatives will help to bring legal coherence between the two human rights systems.

The following sections explore the HRDD responsibilities of online platforms moderating illegal content online, clarifying the specific responsibilities applicable to the moderation of the most serious cases of hate speech.[72] The focus on criminal hate speech provides a clear and more foreseeable legal basis for the extrapolation of corporate HRDD responsibilities.

3.3 Broader framework: AI and the corporate responsibility to respect human rights

This section presents key international and European HRDD instruments relevant to AI businesses, such as online platforms, moderating content online: the UNGPs, OECD initiatives, the CSDDD and the AI Act. The main takeaway is that all AI businesses have the responsibility to adopt a *policy commitment* to respect human rights. Applied to online platforms, this can be interpreted to mean that they should explain in their ToS how their content moderation respect human rights.

3.3.1 United Nations Guiding Principles on Business and Human Rights

Unanimously endorsed in 2011 by the United Nations Human Rights Council, the UNGPs articulate a universal framework for the prevention and mitigation of human rights interference by businesses.[73] Though not legally binding, the UNGPs constitute the most influential international expression of the

71 European Commission, 'The Commission proposes to extend the list of "EU crimes" to hate speech and hate crime' (n 23).

72 Tarlach McGonagle, 'Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation', in Oxford Handbooks in Law; The Oxford Handbook of Online Intermediary Liability (2020), available at <https://doi:10.1093/oxfordhb/9780198837138.013.24> accessed 19 November 2024, 15.

73 Previous attempts had failed, including the UN Sub Commission on the Promotion and Protection of Human Rights' 'Draft Norms on the Responsibilities of Transnational Corporations and other Business Enterprises with regard to Human Rights' (2003) E/CN.4/Sub.2/2003/12/Rev.2, which foresaw binding responsibilities for business enterprises. See Nicola Jägers, 'UN Guiding Principles on Business and Human Rights: Making headway towards real corporate accountability?' (2011) 29(2) Netherlands Quarterly of Human Rights 159–163, available at <https://doi.org/10.1177/016934411 102900201> accessed 19 November 2024, cited in Lottie Lane, 'Artificial Intelligence and Human Rights: Corporate Responsibility Under International Human Rights Law', in Aleš Završnik and Katja Simonèiè (eds), *Artificial Intelligence, Social Harms and Human Rights* (Palgrave Macmillan 2023) 183-205, 188. See also Joseph and Kyriakakis (n 44).

corporate responsibility to respect human rights, particularly through HRDD.[74] The UNGPs clarify that businesses should have in place *policies* and *processes* to respect human rights including: (a) a policy commitment to respect human rights; (b) a HRDD process to identify, prevent, mitigate and account for adverse impacts on human rights; (c) processes to enable the remediation of any human rights abuses.[75]

Applying Principle 15 to online platforms, the policy commitment can arguably be reflected in a more detailed manner in a company's ToS. Typically, ToS are legal agreements between online platforms and their users containing, among other topics, the allowed/prohibited online conduct and explaining how the company considers human rights.[76] ToS guide the machine learning and large language models used to moderate content online. As the main publicly available policy tool used by online platforms to communicate with their users the rules guiding their services applicable to both users and the platform itself, ToS can be said to fulfil the purpose of the corporate policy commitment to respect human rights.

The HRDD commitment is essential to identify, prevent, mitigate and account for actual and potential human rights abuses by businesses.[77] Notably, the UNGPs prescribe that HRDD should involve meaningful consultation with potentially affected groups. Applying this conceptualisation of HRDD to online platforms, these arguably have the HRDD responsibility to better engage with people targeted by harmful content hosted by them. One way could be by employing a community-driven contextualisation of hate speech (applicable to cases of hate speech not criminally actionable) in ToS. Further, this Chapter proposes that preventive HRDD responsibilities requires online platforms to revisit their policy commitments to adequately reflect their commitments to human rights, hence the HRDD responsibility to review existing ToS.

Reflecting the commitment to respect human rights in ToS is all the more important given the non-binding nature of the UNGPs and would be a complementary measure to clarify the applicability of the existing human rights regulatory and policy frameworks to corporations. Furthermore, the rising

74 Robert McCorquodale and Justine Nolan, 'The Effectiveness of Human Rights Due Diligence for Preventing Business Human Rights Abuses' (2021) 68 Netherlands International Law Review 455, cited in Lottie Lane, 'Preventing long-term risks to human rights in smart cities: a critical review of responsibilities for private developers of AI' (2023) 12(1) Internet Policy Review. See also Joseph and Kyriakakis (n 44). The UNGPs details the State duty to protect human rights and victims' access to remedy resulting from corporate abuses, respectively in Pillars 1 and 3. See Surya Deva, 'Guiding Principles on Business and Human Rights: Implications for Companies' (2012) 9(2) European Company Law 101.

75 UNGPs (n 13), Principle 15; Lottie Lane, 'A Human Rights Responsibility Primer for Businesses Developing AI: Part 2' (*Medium*, 14 September 2021), available at <https://medium.com/slimmerai/a-human-rights-responsibility-primer-for-businesses-developing-ai-part-3-68f1e5b33e20> accessed 6 April 2023.

76 UNGPs (n 13), Principle 16.

77 UNGPs (n 13), Principle 17.

debate about services provided by large online platforms possibly amounting to public services essential for the exercise of the right to freedom of expression,[78] strengthens the argument that the businesses' freedom to decide what to include in ToS should be proportionally restricted to give primacy to HRDD and to reflecting human rights standards in ToS.[79]

3.3.2 Initiatives by the Organization for Economic Cooperation and Development

The OECD has also developed numerous initiatives that shed light on the corporate HRDD framework. First, the OECD Declaration and Guidelines for Multinational Enterprises comprising recommendations to conduct responsible business were first adopted in 1976 and updated in 2011 to include a chapter on human rights in line with the UNGPs.[80] In 2018, the OECD adopted its Due Diligence Guidance for Responsible Business Conduct[81] to help companies implement the Guidelines for Multinational Enterprises and understand the application of due diligence principles. This Guidance refers to the UNGPs and suggests that HRDD includes the elements demonstrated in Figure 2 below.

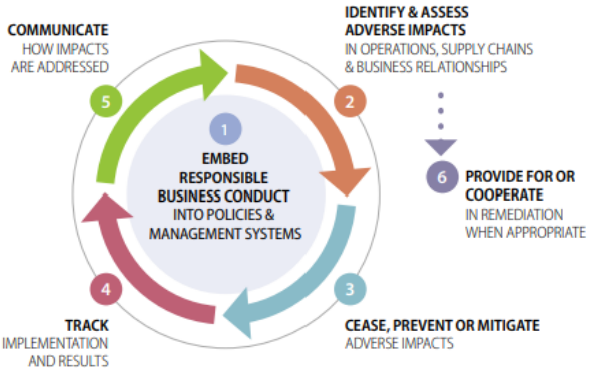

Figure 2 – OECD Due diligence process and supporting measures[82]

78 See above (n 12). Additionally, the argument that online platforms provide services reaching a general public has been reiterated in European judicial instances, e.g. ECLI:EU:C:2021:503, paragraph 69.

79 Interestingly, in a recent opinion, Advocate General Rantos of the European Court of Justice (ECJ) broadly noted that online platforms must not design or enforce ToS contrary to EU law. See ECLI:EU:C:2022:704, Opinion of Advocate General Rantos delivered on 20 September 2022 (1) case C-252/21, Paragraph 78.

80 OECD, 'OECD Guidelines for Multinational Enterprises' (2011), available at <http://mneguidelines.oecd.org/guidelines/> accessed 6 April 2023.

81 OECD, 'OECD Due Diligence Guidance for Responsible Business Conduct' (2018), available at <https://www.oecd.org/investment/due-diligence-guidance-for-responsible-business-conduct.htm > accessed 6 April 2023.

82 OECD (n 81), 21.

Chapter 4 of the 2018 Guidance expands on HRDD and reiterates the importance that businesses embed human rights into their policies. To do this, the Guidance suggests that businesses make the commitment publicly available, for example on their website, underlining its importance to business relationships.[83] The Guidance also explains that consumers or end-users of products, as persons or groups whose interests can be affected by the companies' activities, must be informed about the due diligence processes shaping the companies' operations.[84] Applying this framework to online platforms, it can be argued that, though not in a legally binding way, ToS communicated by online platforms to their users are typically the tool fulfilling the purpose of the OECD's policy commitment standard.

In recent years, the OECD has adopted instruments specifically addressing HRDD for AI businesses, and thus with impact for online platforms. The Recommendation of the Council on AI, adopted in 2019,[85] stresses that AI businesses should respect the rule of law, human rights and democratic values throughout the AI system lifecycle, including the right to non-discrimination.[86] AI actors should also commit to transparency and explainability to promote a better understanding of the AI systems and to enable stakeholders to understand the outcome of decisions led by AI systems.[87] Applying this framework to online platforms, it can similarly be argued that ToS constitute an adequate tool for AI actors to communicate to their users in a transparent and explainable manner their automated-decision making algorithms used for large scale content moderation.

Finally, in 2021, the OECD published its annual publication Business and Finance Outlook 2021: AI in Business and Finance.[88] The potential contribution of automated content moderation to the proliferation of illegal content online is expressly raised as a key concern and it is suggested that content moderation policies balance freedom of expression with general human rights safeguards (e.g. right to appeal and to remedy).[89] This instrument reiterates two essential

83 OECD (n 81), Ibid 22.

84 OECD (n 81), 48.

85 OECD, Recommendation of the Council on Artificial Intelligence (2019), available at <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#mainText> accessed 6 April 2023.

86 OECD (n 85), Paragraph 1.2.

87 OECD (n 85), Paragraph 1.3. The terms 'transparency' and 'explainability' are defined in multiple ways in AI governance initiatives. Transparency typically concerns access to information regarding when/how AI systems are used, but also of HRDD processes. Explainability most often refers to making available information regarding how systems reach their outcomes in a way that is accessible and understandable to end-users. For a more in-depth discussion, see Lottie Lane, 'Artificial Intelligence and Human Rights: Corporate responsibility in AI governance initiatives' (2023) Nordic Journal of Human Rights, available at <https://doi.org/10.1080/18918131.2022.2137288> accessed 19 November 2024.

88 OECD, 'Business and Finance Outlook 2021', available at <https://www.oecd.org/finance/oecd-business-and-finance-outlook-26172577.htm> accessed 6 April 2023.

89 OECD (n 88), 3.2.4.

points: the need to implement HRDD throughout the whole cycle of business operations and the need to develop explainable AI systems.[90]

3.3.3 EU Directive on Corporate Sustainability Due Diligence

The UN and OECD initiatives are key to introducing HRDD and have significantly influenced the legislative framework on HRDD under development in the EU.[91] In June 2023, the European Parliament adopted a draft detailing many amendments to the CSDDD, which was first proposed by the European Commission in February 2022.[92] In June 2024, the European Parliament and the Council adopted the CSDDD.[93]

This Directive will be enforced by national authorities and by a European Network of Supervisory Authorities to be set up by the Commission. Although the scope of the CSDDD has been broadened during the negotiations, it remains more limited than the UNGPs and the OECD's guidance, covering EU companies with 250+ employees and a turnover of over _40 million worldwide and non-EU companies with an equivalent turnover threshold generated in the EU.[94] With regard to companies with lower revenue and fewer employees, the CSDDD extends due diligence responsibilities for some companies operating in 'high-impact sectors', which do not currently include online platforms.[95]

The CSDDD requires relevant companies to: 1) integrate HRDD into policies and management systems; 2) identify and assess adverse human rights and environment impacts; 3) prevent, cease and minimise adverse human rights and environment impacts; 4) assess the effectiveness of measures; 4) communi-

90 Lane, 'Preventing long-term risks to human rights in smart cities' (n 74) 3.1.

91 This is evident in, e.g., European Parliament, 'Amendments adopted by the European Parliament on 1 June 2023 on the proposal for a directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937' COM(2022)0071 – C9-0050/2022 – 2022/0051(COD) Recitals 5, 6, 12, 16, 22.

92 Ibid and CSDDD (n 15), respectively. As a Directive, if adopted, the CSDDD would be legally binding on EU countries, setting a goal to be attained at the EU level whilst giving individual countries the freedom to decide which laws to adopt to reach such a goal. The move towards mandatory HRDD at the EU level is part of a broader movement towards binding HRDD standards for private companies, including at the international and national level. For critical discussion, see e.g. Sarah Joseph and Joana Kyriakakis (44); Chiara Macchi and Claire Bright, 'Hardening Soft Law: The Implementation of Human Rights Due Diligence Requirements in Domestic Legislation' in Martina Buscemi et al (eds) Legal Sources in Business and Human Rights : Evolving Dynamics in International and European Law (Brill 2020) 218; Surya Deva, 'Mandatory human rights due diligence laws in Europe: A mirage for rightsholders?' (2023) 36(2) LJIL 389.

93 CSDDD (n 15).

94 The CSDDD General Approach had a narrower scope, applying to EU companies with 500+ employees and an annual turnover of _150 million: 'Article 1.

95 Ibid, Recitals 21-23, 15.

cate; and 6) provide remedial mechanisms for human rights and environmental negative impacts caused by their own operations, their subsidiaries and their value chains.[96] This places an important emphasis on 'preventive responsibilities'[97] to mitigate or avoid potential harms rather than only taking action once harm has already occurred.[98]

Importantly, regarding the CSDDD's conceptualisation of human rights, Annex I focuses on international human rights law, excluding regional European human rights law (i.e. the ECHR and the CFREU). The CSDDD could have effects beyond EU companies in some situations.[99] However, omitting references to key and legally binding European instruments protecting human rights applicable to all Member States (MS) of the EU whilst referring to non-binding, international standards and international treaties that have not been universally adopted has been criticised.[100]

Applying this framework to the regulation of hate speech in Europe can lead to legal incoherence because the most concrete attempt to conceptualise hate speech and its most serious forms was developed at the CoE level in CM/Rec(2022)16 (explained above in Section 3.2.2.2). This potential legal incoherence could be addressed preemptively because, although the list in Annex I is restricted to international human rights law, international human rights such as the right to life, liberty and security,[101] the prohibition of inhuman or degrading treatment,[102] the prohibition of discrimination[103] can be interpreted to include the conceptualisation of criminal hate speech as per CM/Rec(2022)16.

Part I of Annex I expressly acknowledges the application of *inter alia* Article 7 ICCPR and, part II of the same Annex expressly acknowledges *inter alia* the UDHR and the ICCPR.[104] Thus, for online platforms falling under its scope, the CSDDD's provisions could be applicable to online moderation of hate speech. Proposed Recital 22 of the CSDDD reflects this, mentioning that 'the Commission should develop sector-specific guidelines', including in

96 CSDDD (n 15) 32 (16). Articles 4-11.
97 McCorquodale (n 74) cited in Lane, 'Preventing long-term risks to human rights in smart cities' (n 74).
98 See especially Arts. 6, 7 and 10.
99 Its application to non-EU companies falling under the scope of Article 1 CSDDD, as well as the so-called 'Brussels effect' of EU regulation across the globe. See Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (OUP 2020).
100 Claire Methven O'Brien and Jaques Hartmann, 'The European Commission's proposal for a Directive on corporate sustainability due diligence: two paradoxes' (*EJIL: Talk!*, 19 May 2022), available at <https://www.ejiltalk.org/the-european-commissions-proposal-for-a-directive-on-corporate-sustainability-due-diligence-two-paradoxes/> accessed 6 April 2023.
101 Universal Declaration of Human Rights (UDHR) Article 3; International Covenant on Civil and Political Rights (ICCPR) Article 6.
102 UDHR Article 5, ICCPR Article 7.
103 UDHR Article 7, ICCPR Article 4.
104 CSDDD (n 15).

relation to 'the production, provision and distribution of information and communication technologies or related services, including … artificial intelligence, … social media and networking … and other platform services'.[105] Interestingly, these are not mentioned in Article 13(1)(a)(c), which suggests specific sectors for which guidelines should be adopted. Nevertheless, online platforms could fall within the scope of Article 13, which is not phrased as constituting an exhaustive list.

3.3.4 EU Artificial Intelligence Act

The EU initiated a legislative process to regulate the responsibilities of AI businesses when, in April 2021, the EC proposed a Regulation on harmonised rules on artificial intelligence (AI Act).[106] In December 2022, the Council adopted its General Approach to the AI Act[107] and, in June 2023, the EP adopted its Draft Compromise Amendments proposed by the Committee on Internal Market and Consumer Protection (IMCO) and by the Committee on Civil Liberties, Justice and Home Affairs (LIBE).[108] In June 2024, the EP and the Council of the EU adopted the AI Act , which entered into force in August 2024.[109]

Though not framed as a human rights instrument, one of the AI Act's objectives is for AI systems to 'ensure a high level of protection of … fundamental rights … from harmful effects of artificial intelligence systems in the Union'.[110] AI systems are defined as 'machine-based system that is designed

105 European Parliament 2022/0051(COD) on the CSDDD (n 91).

106 European Commission, 'Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, COM(2021) 206 final.

107 Council of the European Union, 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General approach'. Interinstitutional File 2021/0106(COD).

108 European Parliament, Compromise Amendments on the Draft Report Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD)), adopted 14 June 2023, available at <https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf> accessed 19 November 2024.

109 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.

110 Ibid Recital 1; Lottie Lane, 'Clarifying Human Rights Standards through Artificial Intelligence Initiatives: A multi-level comparative analysis' (2022) International and Comparative Law Quarterly 74(1) 16, available at <https://doi.org/10.1017/S0020589322000380> accessed 19 November 2024. However, the scope of human rights protection afforded by the AI Act has been criticised. See e.g. European Digital Rights, 'EU Parliament calls for ban of public

to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments'.[111]

The AI Act takes a two-prong approach of prohibiting certain systems whilst regulating others on the basis of a risk-based approach with three levels of risk posed by AI systems: i) unacceptable risk AI; ii) high-risk AI; iii) low or minimal risk AI.

The potential application of the AI Act to content moderation by online platforms can be summarised in two ways. First, AI systems used for content moderation systems can be prohibited under Article 5(1)(a) of the AI Act. This provision prohibits systems that 'deplo[y] subliminal techniques beyond a person's consciousness with the objective to or the effect of materially distorting a person's behaviour in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm'. According to Amnesty International, this was arguably the case when Meta (formerly Facebook) did not remove and even amplified hate speech towards the Rohingya Muslims in Myanmar, potentially contributing to adverse impact on their human rights.[112]

Second, for AI systems falling outside the remit of Article 5 of the AI Act, recommender systems in content moderation that impact the administration of justice and democratic processes may end up being considered high-risk AI systems, should the EP's amendments be adopted in the final text. Article 6(3) defines high-risk systems as those whose output is not 'purely accessory in respect of the relevant action or decision to be taken and is not therefore likely to lead to a significant risk to the health, safety or fundamental rights'.[113] This is arguably the case for content moderation systems that either allow or promote material containing hate speech, which, due to the vast number of posts to be monitored on social media platforms, are subject to minimal human oversight, making the systems more than 'purely accessory' to decisions to remove content.

Content moderation AI systems could also be considered to fall indirectly under the scope of 'high-risk' systems in limited situations. As explained in Section 3.4.1 below, the DSA prescribes the HRDD responsibility for social

---

facial recognition, but leaves human rights gaps in final position on AI Act' (14 June 2023), available at <https://edri.org/our-work/eu-parliament-plenary-ban-of-public-facial-recognition-human-rights-gaps-ai-act/> accessed 18 November 2024.

111 AI Act (n 109) Article 3.

112 Amnesty International, 'Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya' (2022), available at <https://www.amnesty.org/en/documents/asa16/5933/2022/en/> accessed 19 November 2024, and Al Jazeera, 'Rohingya sue Facebook for $150bn for fuelling Myanmar hate speech' (n 9).

113 Lane, 'Preventing long-term risks to human rights in smart cities' (n 74).

media companies to report criminal offences, including criminal hate speech, to law enforcement. When law enforcement agencies use the results of content moderation AI systems to assess the 'risk for a natural person to become a potential victim of criminal offences', which could include physical or psychological harm caused by hate speech, Arguably, this AI system falls under the scope of Paragraph 6(a) of Annex III, which labels such systems as high risk.

Should these provisions apply, providers[114] of content moderation systems would be subject to a number of risk-management standards reflecting various elements of HRDD. This includes an obligation to identify and analyse 'known and foreseeable risks most likely to occur…in view of the intended purpose of the system'.[115] Further, Article 9(2)(d) requires providers of high-risk systems to adopt 'suitable risk management measures' to respond to risks. Arguably, this could include a prohibition of criminal hate speech in the ToS of platforms in relation to their content moderation systems.

3.4 Specific framework: preventive corporate responsibilities to counter online hate speech

This section examines how preventive HRDD responsibilities apply to the moderation of criminal hate speech in Europe. The growing clarity regarding the European conceptualisation of the criminal hate speech, namely with the adoption of CM/Rec(2022)16, enables a common regional understanding of criminal hate speech from which specific HRDD responsibilities can be developed. The main instruments analysed are the DSA, the AVMSD, the EU Code of conduct and Recommendation CM/Rec(2022)16.

While Section 3.3. clarified that AI businesses, including online platforms, must adopt a *policy commitment* to respect human rights, this section expands on the preventive HRDD responsibilities. We explain that the current European regulatory system suggests that online platforms should reflect their commitment to human rights in the ToS, including to human rights standards applicable to counter online hate speech, such as the right to respect for private and family life and the prohibition of discrimination.

The EU has developed frameworks regulating content moderation and the HRDD responsibility of online platforms not to host illegal online content, such as hate speech. Video-sharing platforms have the heightened responsibility to explicitly reflect the prohibition of hate speech in their ToS. While such HRDD frameworks take different approaches, they reflect the importance of

114 Article 3(2) defines a 'provider' as 'a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed and places that system on the market or puts it into service under its own name or trademark, whether for payment or free of charge'. This would cover online platforms.

115 Article 9(2)(a); Lane, 'Preventing long-term risks to human rights in smart cities' (n 74).

including the prohibited content in the ToS, which should be conveyed to users in a clear and transparent manner.

3.4.1 EU Regulation on a Single Market for Digital Services

The most relevant instrument at the EU level articulating HRDD responsibilities for online platforms is the Regulation on a Single Market For Digital Services (DSA), in force since November 2022.[116] The DSA regulates online platforms operating in an online environment under the EU territorial jurisdiction and establishes HRDD responsibilities for different online stakeholders, depending on their role, size and impact.[117]

Noting that some of the biggest online platforms are based in the USA, the DSA introduces a new regulatory approach as, aside from having to comply with the legal framework in the USA, companies now also have to adapt to European legal standards in operations conducted in the EU. For example, the legal framework on freedom of expression is significantly distinct as the USA gives primacy to the First Amendment whereas in the EU, though freedom of expression is considered a quintessential human right in democratic societies, the conditions for restrictions in cases of discrimination are expressly prescribed.[118]

The DSA sets due diligence responsibilities (Chapter III) for various stakeholders, including for online platforms (Sections 2, 3 and 4) and for Very Large Online Platforms (VLOPs) (Section 5).[119] The DSA provides general instructions to intermediary services to 'diligently regard' fundamental rights of the users as enshrined in the CFREU.[120] This Chapter examines the HRDD rules in the DSA applicable to online platforms, with a particular focus on

116 DSA (n 17). The DSA is a legal instrument in the form of a EU Regulation and directly regulates the means through which MS must achieve the prescribed goals.

117 DSA, Recital 41.

118 LICRA v Yahoo! and Association "Union des Etudiants Juifs de France", la "Ligue contre le Racisme et l'Antisémitisme", le "MRAP" (intervenant volontaire) / Yahoo ! Inc. et Yahoo France, available at <https://www.iddn.org/cgi-iddn/french/affiche-jnet.cgi?droite=decisions/responsabilite/ord_tgi-paris_201100.htm> accessed 19 November 2024. For a comparison of the USA and European legal frameworks on freedom of expression, see e.g. Brittan Heller and Joris van Hoboken, 'Freedom of Expression: A Comparative Summary of United States and European Law' (2019) Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression Working Paper, available at <https://www.ivir.nl/publicaties/download/TWG_Freedom_of_Expression.pdf> accessed 19 November 2024.

119 Provisions applicable to all providers of intermediary services are covered in Section I of the DSA.

120 DSA, Article 14(4). For a more detailed analysis on Article 14 DSA see Naomi Appelman, João Pedro Quintais and Ronan Fahy, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation' (2023) German Law Journal.

the HRDD responsibilities of VLOPs as these represent the category of stakeholders posing higher risks of disseminating illegal content.

The DSA introduces asymmetric HRDD for VLOPs precisely because of their far reach, high turnover and their ability to comply with stronger HRDD requirements.[121] The heightened HRDD threshold for VLOPs is also reflected in Article 34 of the DSA which states that VLOPs 'shall identify, analyse and assess (…) any significant systemic risks' which include: (a) the dissemination of illegal content through their services; and (b) any negative effect for the exercise of the fundamental rights to respect for private and family life, freedom of expression and information, the prohibition of discrimination and the rights of the child, as enshrined in Articles 7, 11, 21 and 24 of the Charter, respectively.'[122]

The DSA specifically mentions that the dissemination of hate speech pertains to the first category of systemic risks to be assessed by VLOPs[123] and that it falls under the category of illegal content in the EU.[124] In fact, in the Explanatory memorandum, it is clarified that the DSA builds on the Recommendation on illegal content of 2018,[125] which already mentioned hate speech as illegal content in the EU. Furthermore, the DSA confirms to build upon self-regulatory initiatives such as codes of conduct or other self-regulatory measures which, in the framework applicable to hate speech, includes the Code of conduct against illegal hate speech.[126]

A concrete proposal in the DSA for VLOPs to mitigate systemic risks, such as the dissemination of hate speech, is by 'clearly and unambiguously' informing users of ToS as well as remedies and redress mechanisms,[127] adapting their terms and conditions and their enforcement,[128] and

> 'adapting their content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision making processes and dedicated resources for content moderation'.[129]

The inclusion of ToS in Article 35 of the DSA as a mitigatory measure must be interpreted as a recognition that ToS are a key tool to provide legal clarity,

121 DSA, 11.
122 DSA, Article 34.
123 DSA, Recital 80.
124 DSA, Recital 12.
125 EC, Recommendation (EU) 2018/334 (n 19).
126 DSA, Recital 88.
127 DSA, Article 14(5).
128 DSA, Article 35(1)(b)
129 DSA, Article 35(1)(c).

foreseeability and transparency to users of VLOPs. As such, ToS represent the ideal communication tool to contain the conceptualisation of what is illegal content, and hence, what is hate speech.

The DSA does not indicate to VLOPs the conceptualisation of illegal content that they should follow in their ToS. We suggest that the conceptualisation of hate speech should be limited to the most serious forms of hate speech, i.e. criminal hate speech. The limitation of the requirement to reflect the conceptualisation of hate speech to criminal hate speech is justified by the European human rights understanding in CM/Rec/(2022)16. The EC proposed to add hate speech to the list of EU crimes[130] but, in the meantime and while the EU does not conceptualise the main elements of criminal hate, CM/Rec/(2022) 16 summarises the European human rights standards for the conceptualisation of the most serious forms of hate speech.

Additionally, VLOPs should inform their users through their ToS that such criminal offences are reported to competent law enforcement authorities. The DSA prescribes the due diligence measure applicable to all providers of hosting services, including online platforms, that criminal offences involving a threat to life be reported to law enforcement or judicial authorities.[131] Given the additional requirements for VLOPs regarding information and transparency of their ToS, also in the context of cooperation with law enforcement, businesses should utilise ToS to clearly inform their users of the companies' HRDD responsibilities.

Applying this framework to countering online hate speech, it is possible to propose minimum best practices for the improvement of legal clarity and coherence in European human rights frameworks – VLOPs should explicitly mention in their ToS the minimum European human rights elements in the conceptualisation of criminal hate speech and inform users that speech is removed and reported to law enforcement.

Should the EC add hate speech to the list of EU crimes,[132] it becomes imperative for VLOPs to explain to users in their ToS what the framework of criminal hate speech is and what consequences it bears for users posting such illegal content i.e. referral to law enforcement. Under Article 35(3), the Commission could issue guidelines on best practices since the abovementioned legal avenues would support the businesses' compliance with transparency and clarity requirements regarding ToS in the DSA but also more generally with the businesses' preventive corporate HRDD responsibilities.

130 European Commission, 'The Commission proposes to extend the list of "EU crimes" to hate speech and hate crime' (n 23).

131 DSA, Article 18.

132 European Commission, 'The Commission proposes to extend the list of "EU crimes" to hate speech and hate crime' (n 23).

3.4.2 EU Audiovisual Media Services Directive

Another relevant EU legal instrument creating HRDD responsibilities for online platforms is the 2018 revised AVMSD.[133] The AVMSD regulates the activity of TV broadcasters, video-on-demand services and video-sharing platforms. With particular relevance for this Chapter, video-sharing platforms include commercial services devoted to making available to the general public programmes and user-generated videos with the purpose to inform, entertain or educate, shared via the Internet, and where content organisation is determined by the video-sharing platform (i.e. the service displays, tag and recommends video content to the users).[134]

While the AVMSD mostly imposes obligations on Member States in their regulation of audiovisual media services, it also directly establishes minimum standards to be adhered to by businesses.[135] Regarding the prohibited content, the AVMSD initially refers to the conceptualisation of hate speech in Council Framework Decision 2008/913/JHA, which is limited to acts of racism and xenophobia.[136] Nevertheless, similarly to the DSA, the AVMSD expressly prohibits disseminating 'incitement to violence or hatred directed against a group of persons or a member of a group based on any of the grounds referred to in Article 21 of the [CFREU]'. Further, the AVMSD recognises that this reference to the Council Framework Decision should be applied 'to the appropriate extent', and Articles 6(a)(a) and 28(1)(b) AVMSD refer to an expansive interpretation of protected categories in Article 21 CFREU.[137] It is therefore possible to argue that the AVMSD considers a broader, intersectional conceptualisation of hate speech.

The AVMSD introduces substantial developments concerning the responsibility framework for video-sharing platforms, including specific measures to comply with the prohibition to host hate speech. Article 28(3) prescribes that video-sharing platform services *shall* implement compliance measures *consisting of* including and applying in the terms and conditions the requirements in Articles 28(1) and 9(1). While not specifically phrased as due diligence measures, Article 28(3)(i) and (ii) AVMSD symbolise a milestone in HRDD as they expressly address the considerable role of ToS in ensuring a clear and transparent tool for a public commitment to respect human rights.

---

133 AVMSD (n 18). AVMSD Recital 45 introduces obligations for Member States to ensure that audiovisual media services increase protection of minors from harmful content and protection of the general public from hate speech.

134 AVMSD, Article 1(1)(b)(aa).

135 E.g. AVMSD, Article 28.

136 AVMSD, Recital 17.

137 AVMSD, Recital 17. Furthermore, Article 9(1)(c) AVMSD stipulates that 'audiovisual commercial communications shall not (i) prejudice respect for human dignity; [or] (ii) include or promote any discrimination based on sex, racial or ethnic origin, nationality, religion or belief, disability, age or sexual orientation'.

3.4.3 EU Code of conduct on countering illegal hate speech online

In 2016, the EU agreed with some of the largest online platforms on a 'Code of conduct on countering illegal hate speech online' (CoC); of note, Facebook, Twitter and YouTube were among the first signatories.[138] The CoC is a co-regulatory instrument which, though resulting in legal consequences if breached, is arguably difficult to enforce. However, it represents a strong acknowledgement of the rise of hatred in the digital environment and it likewise symbolises a strong policy commitment from online platforms to counter online hate speech. The very restrictive conceptualisation of hate speech in the CoC was already criticised in Section 3.2.2. of this Chapter, particularly in comparison to Article 21 of the CFREU and the AVMSD. We adopt a broader intersectional interpretation of the impermissible grounds for hate speech.

For what concerns the HRDD responsibilities imposed upon such online platforms to counter 'incitement to violence and hateful conduct', the CoC points directly to the relevance of containing 'clear information on individual company Rules and Community Guidelines' as a means to improve notices and flagging of said content.[139] This requirement directly follows the preventive HRDD responsibility to commit to respect human rights in a policy statement and during operations.

Moreover, online platforms signatories to this CoC are required to 'educate and raise awareness with their users about the types of content not permitted under their rules and community guidelines', which is a preventive responsibility to help mitigate and avoid potential risks. Additionally, these businesses must have in place 'clear and effective processes to review notifications', which could function as a more responsive measure to risks that may already have incentivise, depending on the actual content of the flagged material.[140] These requirements align with the corporate human rights responsibility to respond to risks by having HRDD procedures in place to identify, prevent, mitigate and account for human rights abuses.

138 CoC (n 20).

139 CoC (n 20), 2.

140 CoC (n 20), 2. Aside from these main HRDD responsibilities, signatories to the CoC must also comply with additional responsibilities such as ensuring that there are civil society organisations fulfilling the role of 'trusted flaggers' and providing regular training on hate speech policies to their staff, 3.

3.4.4 Council of Europe Committee of Ministers' Recommendation CM/Rec(2022)16

Adopted in May 2022, the Recommendation CM/Rec(2022)16 is a milestone achievement in combating online hate speech in Europe. Though not legally binding, CM/Rec(2022)16 represents a clear political pledge of the statutory decision-making body of the CoE: the Committee of Ministers. CM/Rec(2022)16 articulates concrete guidance to all 46 Member States for a comprehensive human rights framework to address hate speech, including in the digital sphere. This means that the guidance provided is inevitably also addressed to the 27 EU Member States. Furthermore, the EU human rights standards draw inspiration from those at the Council of Europe.[141]

Paragraph 31 of CM/Rec(2022)16 clearly provides that:

> 'internet intermediaries should ensure that human rights and standards guide their content moderation policies and practices with regard to hate speech, *explicitly state that in their terms of service* and ensure the greatest possible transparency with regard to those policies, including the mechanisms and criteria for content moderation'.[142]

This standard is complemented by CM/Rec(2018)2 on the roles and responsibilities of internet intermediaries,[143] which underlines that internet intermediaries are responsible for respecting human rights and for implementing adequate measures to that end.[144] It adds that intermediaries whose services pose a higher risk of potential adverse impacts on human rights should adopt greater precautionary measures. Again, one example of such precautionary measures is the careful development and application of the ToS. Moreover, CM/Rec(2018)2 stresses the importance of drafting and applying ToS agreements, community standards and content-restriction policies in a transparent fashion.[145] Nevertheless, companies must still comply with their HRDD responsibilities throughout their operations, i.e. the design, development and deployment of content moderation systems.

Finally, CM/Rec(2022)16 also contains significant advances concerning the responsibilities of internet intermediaries to moderate criminal hate speech. Paragraph 32 articulates the responsibility of internet intermediaries to remove only the most severe cases of hate speech, i.e. criminal hate speech. This appears to be more reactive than preventive and is not expressly referred to

141 Even in the CoC, the parties refer to the jurisprudence of the ECtHR. See footnote 1 of the CoC.

142 Emphasis added.

143 Recommendation CM/Rec (2018)2 of the CM of the CoE to member States on the roles and responsibilities of internet intermediaries.

144 Ibid paragraph 2.1.2.

145 Ibid paragraph 2.2.2. This can be accomplished, for example, by starting to involve human rights experts in the drafting of ToS.

as a corporate HRDD responsibility. However, it does reflect the HRDD responsibilities found in the UNGPs and the OECD's guidance to take measures to cease or mitigate adverse impacts.[146] Similarly, the suggestion in Paragraph 2.2. that intermediaries report instances of criminal hate speech to public authorities reflects the HRDD measure to provide for or cooperate in remediation when appropriate.[147]

Furthermore, Paragraph 18 CM/Rec(2022)16 clarifies that intermediaries are 'to respect human rights, including the legislation on hate speech, to apply the principles of human rights due diligence throughout their operations and policies, and to take measures in line with existing frameworks and procedures to combat hate speech'.[148] CM/Rec(2022)16 goes beyond the responsibility to remove criminal hate speech to include that

> 'internet intermediaries, including social media, should review their online advertising systems and the use of micro-targeting, *content amplification and recommendation systems* and the underlying data-collection strategies to ensure that they do not, directly or indirectly, *promote or incentivise the dissemination of hate speech*.'[149]

This invites corporations to conduct a full revision of their business models to ensure that their content moderation algorithms are specifically designed to not disseminate, recommend or profit from hate speech.

3.4.5 Proposal of a legal standard

The analysis of the European regulatory and policy human rights framework on criminal hate speech (Section 3.2) and on the corporate preventive HRDD responsibilities to respect human rights (Section 3.3) and to counter hate speech (Section 3.4) suggests that online platforms should reflect as a minimum legal standard the most serious cases of hate speech, i.e. *criminal hate speech*, in their ToS.. To clarify, this Chapter does not take the position that the existing legal framework fails to regulate criminal hate speech on online platforms. Instead, this Chapter claims that the human rights framework on criminal hate speech and the HRDD framework applicable to online platforms do not directly explain their implications for the drafting of the ToS of online platforms.

146 OECD 2018 HRDD guidance (n 81); UNGPs (n 13); Section 3 of this Chapter.
147 Ibid.
148 Regarding the responsibilities of intermediaries moderating hate speech prohibited under civil or administrative law, though the default approach for the responsibility of removal advocated in CM/Rec(2022)16 is strictly invoked in cases of criminal hate speech, CM/Rec(2022)16 prescribes that intermediaries deprioritise and contextualise (paragraph 22) and publish transparent reports with disaggregated and comprehensive data on hate speech cases and restrictions (paragraph 25).
149 Ibid paragraph 36 [emphasis added].

Nevertheless, a combined analysis of these human rights frameworks reveals the above explained human rights implications for the drafting of the ToS.

The proposed standard should currently be a recommendation of best practice for the general AI business,[150] but it can be interpreted as a mandatory legal standard specifically for VLOPs, video-sharing platform services and for companies bound by the CSDDD. According to Article 35(3) of the DSA and Article 13 of the CSDDD, the EC could issue guidelines clarifying that VLOPs, video-sharing platforms and platforms falling under the scope of the CSDDD should explicitly mention in their ToS that they prohibit, remove and report criminal hate speech to relevant public authorities. Figure 3 showcases that this legal standard should be implemented in the initial phase of designing policies and management systems of AI business.

The European Commission to issue guidelines clarifying that VLOPs, particularly video-sharing platforms and platforms falling under the scope of the CSDDD, should explicitly mention in ToS that they prohibit, remove and report criminal hate speech to relevant public authorities.

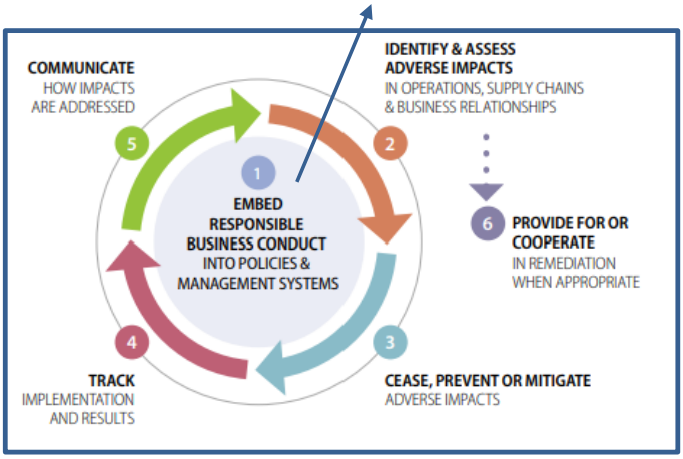

Figure 3 – OECD Due diligence process, including precautionary measures to counter criminal hate speech

150 The recommendation for 'more explicit and specialised guidance' on human rights for AI businesses is stressed by Lane, 'Artificial Intelligence and Human Rights' (n 87).

3.5 Case studies: Compliance of 'Terms of Service' with the conceptualization of criminal hate speech[151]

The previous sections clarified that this Chapter builds upon the conceptualisation of the most serious forms of hate speech (i.e. criminal hate speech) as reflected in the CM/Rec(2022)16 (Section 3.2), as well as the application of corporate HRDD responsibilities of AI businesses to online platforms (Section 3.3), with a particular focus on how companies should conceptualise hate speech in their ToS (Section 3.4). Section 3.5. presents an empirical qualitative analysis of three case studies assessing the compliance of online platforms' ToS with the understanding of the most serious forms of hate speech as prescribed in Paragraph 11 of the CM/Rec(2022)16 (cited in Section 3.2.2.2).[152]

The online platforms studied are Facebook (Section 3.5.1), Twitter (now X)[153] (Section 3.5.2) and YouTube (Section 3.5.3). These platforms were selected based on the following criteria: (1) they fall under the scope of the CSDDD;[154] (2) they are signatories to the CoC; and (3) they qualify as VLOPs as defined in the DSA. As a video-sharing platform, YouTube will also be evaluated in light of the standards in the AVMSD. This section contains additional considerations as to whether the evolving nature of Facebook and even Twitter, as platforms storing and suggesting large amounts of user-generated videos, qualifies them to be assessed in light of the AVMSD.

151 This analysis was conducted in 2023, using data publicly available on the platforms' websites at that time. Changes may apply since platforms frequently update their terms of service. Notwithstanding, the analysis is presented as published in 2023 because it suits the illustrative purpose to show compliance, or lack thereof, of terms of service with he conceptualization of criminal hate speech stemming from European human rights standards.

152 As explained in Section 4.1., though these online platforms are based in the USA and thus typically bound by the USA legal frameworks on freedom of expression, the adoption of most notably the DSA formalises the need of companies operating in the EU territorial jurisdiction to comply with the European regional legal frameworks. For complementary reading, see the European Union Agency for Fundamental Rights "Online Content Moderation – Current Challenges in Detecting Hate Speech" (Vienna, 2023), available at <https://fra.europa.eu/sites/default/files/fra_uploads/fra-2023-online-content-moderation_en.pdf> accessed 26 November 2024.

153 The New York Times, Kate Conger (2023) So What Do We Call Twitter Now Anyway?, available at <https://www.nytimes.com/2023/08/03/technology/twitter-x-tweets-elon-musk.html> accessed 19 November 2024.

154 Though algorithms used by social media are not directly referred to in the General Approach to the AI Act, since the DSA prescribes the HRDD responsibility for social media companies to report criminal offences to law enforcement, this Chapter argues that in such a context, online content moderation systems should be considered a high-risk system, as explained in Section 3.4 above.

3.5.1 Facebook

Facebook is a social media platform owned by Meta Platforms, based in the United States of America (USA). Facebook has close to 3 billion users,[155] with over 250 million active in Europe,[156] and, in October 2022 it was ranked the third most visited website worldwide.[157] As of August 2022, this platform has a turnover of over $100 billion.[158] This company falls under the scope of the CSDDD and qualifies as a VLOP under the DSA. In addition, and with developments witnessed in recent years where Facebook also 'service displays, tags and recommends video content to the users', it arguably also qualifies as a video-sharing platform service under the AVMSD.[159]

Facebook expressly prohibits hate speech in its Community Standards, defining it as 'a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.'[160] It proceeds by providing a definition of attacks as 'violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.'[161] It goes on to expressly prohibit the use of harmful stereotypes which it defines as 'dehumanizing comparisons that have historically been used to attack, intimidate, or exclude specific groups, and that are often linked with offline violence.'[162]

Facebook classifies the severity of hate speech into three 'tiers'.[163] Tier 1 designates the most serious and Tier 3 the least serious forms of hate speech. Tier 1 broadly encompasses hate speech which is violent speech, that establishes dehumanising comparisons or that mocks the concept, events or victims of hate crimes. Tier 2 generally covers hate speech that states inferiority or expressions of contempt, dismissal, disgust, and cursing with the intent to insult. Tier 3 largely encompasses hate speech that excludes or segregates

155 Daniel Ruby, '55+ Facebook Statistics For 2023 (Users, Revenue & Trends)' (*Demand Sage*, 10 February 2023), available at <https://www.demandsage.com/facebook-statistics/> accessed 19 November 2024.

156 Statista, 'Facebook monthly active users (MAU) in Europe as of 4th quarter 2022' (2023), available at <https://www.statista.com/statistics/745400/facebook-europe-mau-by-quarter/> accessed 6 April 2023.

157 Similarweb, Top Websites Ranking (2023), available at <https://www.similarweb.com/top-websites/> accessed 6 April 2023.

158 Mansoor Iqbal, 'Twitter Revenue and Usage Statistics (2023)' (*Business of Apps*, 21 January 2023), available at <https://www.businessofapps.com/data/twitter-statistics/> accessed 19 November 2024.

159 AVMSD, Article 1.

160 Facebook Community Standards Hate speech (2023), available at <https://transparency.fb.com/pt-pt/policies/community-standards/hate-speech/> accessed 19 November 2024.

161 Ibid.

162 Ibid.

163 Ibid.

protected characteristics. These detailed explanations are provided in the Community Standards which are embedded in the website of Meta Transparency Center and could therefore be less accessible to users with less digital literacy. Nevertheless, the prohibition of hate speech is also clearly indicated in Facebook's Help Center, where the prohibition is phrased as 'hate speech, credible threats or direct attacks on an individual or group'.[164] The Help Center is embedded in Facebook's website and directly links to the Meta Transparency Center website where the abovementioned detailed explanation is provided. It is thus possible to conclude that users are well informed about where to access information about Facebook's prohibition of hate speech.

Three key remarks can be made regarding the compliance of Facebook's conceptualisation of hate speech with European human rights standards. First, Facebook does not adopt an open-ended list of protected categories of people, though it recognises that hate speech serves to perpetuate historical oppression.[165] This falls short of the conceptualisation of hate speech that we suggested in Section 3.2. in line with European human rights law, which should explicitly include historically marginalised groups in contexts where hateful expressions are conveyed.[166]

Under Tier 1, Facebook excludes from protection people who 'carried out violent crimes or sexual offenses or representing less than half of a group'. This does not provide an authoritative source dictating whether certain people committed a crime or not. In fact, it was this conceptualisation that enabled Facebook to amplify false allegations that two Muslim tea shop owners had raped a Buddhist girl, fuelling the violence against the Rohingya Muslim community in Myanmar.[167] This clearly violates European human rights standards[168] and should not feature in the company's Community Guidelines.

164 Policies and reporting Legal removal request, What types of things aren't allowed on Facebook? (2023), available at <https://www.facebook.com/help/212826392083694> accessed 19 November 2024.

165 In fact, the disregard for historical oppression and the purely quantitative threshold of considering absolute numbers of people targeted by hate speech as the main metric has led to content moderation decisions violating human rights standards. Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children (2017) available at <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms>.

166 The possibility for gender-based violence on Facebook according to its Community Guidelines was further demonstrated by Audrey Carlsen and Fahima Haque, when they showed that the statement 'Female sports reporters need to be hit in the head with hockey pucks' would not be considered hate speech. The New York Times, What Does Facebook Consider Hate Speech? Take Our Quiz (2017), available at <https://www.nytimes.com/interactive/2017/10/13/technology/facebook-hate-speech-quiz.html> accessed 19 November 2024.

167 Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya (2022), available at <https://www.amnesty.org/en/documents/asa16/5933/2022/en/> accessed 19 November 2024, 25.

168 E.g. Right to an effective remedy and to a fair trial (CFR, Article 47) and Presumption of innocence and right of defence (CFR, Article 48).

Further, if understood as permission for gender-based discrimination, Facebook's acceptance under Tier 2 of 'certain gender-based cursing in a romantic break-up' violates numerous regional and international human rights treaties (e.g. the CFREU,[169] the CoE Convention on preventing and combating violence against women and domestic violence (also known as the Istanbul Convention),[170] the CoE Convention on Cybercrime (also known as the Budapest Convention),[171] and the UN Convention on the Elimination of all Forms of Discrimination Against Women[172]).

Second, from the perspective of the right to freedom of expression, concerning the types of prohibited acts, it is positive to note that Facebook may consider content with intentional satirical tones not to constitute hate speech, as these are often used as counter-speech by people targeted by hate speech.[173] This acknowledgment aligns with the principle set in the ECtHR's *Handyside v United Kingdom* judgement that to support pluralism, tolerance and broadmindedness, 'freedom of expression […] is applicable not only to "information" or "ideas" that are favourably received or regarded as inoffensive or as a matter of indifference but also to those that offend, shock or disturb the State or any sector of the population.'[174]

Third, the classification of hate speech under 3 Tiers fails to align with the conceptualisation of the most serious forms of hate speech as per CM/Rec(2022)16 and the consequences of the classification are not mentioned. Facebook clarifies in its Transparency Center that 'severity' is 'the likelihood that the content could lead to harm both offline and online' and that this is a key factor determining which content the human review teams review first.[175] However, it is not explained whether there is a standard action for review within each Tier. This framework does not align with the standard of differentiating the elements of criminal hate speech as articulated in Paragraph 11 of CM/Rec(2022)16. Moreover, this framework fails to align with Paragraph 32 of CM/Rec(2022)16, requiring that online platforms remove only the most serious forms of hate speech.

Regarding HRDD, Facebook's Transparency Center does include a section dedicated to 'Enforcement' that explains how technology and review teams detect and review potentially violating content and accounts.[176] This com-

169 E.g. Right to human dignity (Article 1); Right to the integrity of the person (Article 3); Respect for private and family life (Article 7).
170 E.g. Psychological violence (Article 33); Stalking (Article 34).
171 E.g. Article 14(2)(b).
172 E.g. Article 2.
173 Facebook (n 160).
174 *Handyside v United Kingdom* App No 5493/72 (07/12/1976), Paragraph 49.
175 How Meta prioritises content for review (2022), available at <https://transparency.fb.com/policies/improving/prioritizing-content-review/>.
176 How Meta enforces its policies (2022), available at <https://transparency.fb.com/en-gb/enforcement/> accessed 19 November 2024.

munication arguably resembles the HRDD phase of identifying and assessing adverse impacts on human rights caused during the business' operations. In addition, Meta also clarifies that it follows a three-part approach to content enforcement encompassing 'removal, reduction and information', reflecting the HRDD responsibilities to cease, prevent, mitigate and communicate adverse impacts on human rights. Nevertheless, while Facebook's enforcement process does resemble the international and regional corporate HRDD standards, there is no formal recognition of such legal inspiration. A formal recognition would be important not only to pay due tribute to the influence of the HRDD in the company's internal enforcement processes but also encourage other platforms to also follow the HRDD framework.

3.5.2 Twitter (now X)[177]

Twitter, Inc. is a social media platform based in the USA, ranked in January 2023 as the fourth most visited website worldwide.[178] This platform has around 1.3 billion accounts,[179] with over 100 million active in Europe.[180] In 2022, it had a turnover of $4.4 billion.[181] This company qualifies as a VLOP as per the DSA and also falls under the CSDDD.

Twitter prohibits 'hateful conduct', which it defines as the promotion or incitement of 'violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease'.[182] Additionally, it prohibits the display of 'hateful imagery and display names', described as 'hateful images or symbols in [a user's] profile image or profile header'.[183]

In assessing the compliance of the conceptualisation of hate speech in Twitter's Community Guidelines with European human rights standards, three aspects are worth discussing. First, Twitter does not articulate an open-ended list of protected categories as it limits the protection of hate speech to people targeted on the basis of the characteristics listed above. This may fail to protect people belonging to a historically oppressed or marginalised group, who should be protected in order to follow an approach promoting legal coherence

177 For the purpose of coherence, this section uses the company name at the time of the study.
178 Similarweb (n 157).
179 Ruby (n 155).
180 Musically, 'YouTube, Meta, Twitter and Spotify (sort of) reveal their EU user figures' (2023), available at <https://musically.com/2023/02/20/youtube-meta-twitter-and-spotify-sort-of-reveal-their-eu-user-figures/> accessed 6 April 2023.
181 Iqbal (n 141).
182 Twitter Help Center, 'Hateful Conduct' (2023), available at <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> accessed 6 April 2023.
183 Ibid.

within European human rights law (see Section 3.5.1 above). Nevertheless, Twitter does acknowledge that its policy to counter hate speech seeks to combat 'abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised', which aligns with the conceptualisation of hate speech by Matsuda.[184] Twitter also allows slurs used between groups if not intended to be hateful and if used as 'a means to reclaim terms that were historically used to demean individuals'.[185] This policy aligns with the standard in CM/Rec(2022)16 on platforms' crucial role in promoting counter-speech and alternative speech.[186]

Second, with respect to the types of prohibited acts, Twitter requires that threats be 'violent' and slurs be 'repeated'.[187] These requirements do not align with the European human rights standards on criminal hate speech. However, the severity (scale of violence) and frequency or reach (scale of relevance) could be interpreted as relevant criteria for the contextualisation of hate speech prohibited under civil or administrative law (as explained in Section 3.2.2.1). Additionally, although Twitter prohibits references to genocides, it does not include a prohibition of denial or trivialisation of other war crimes or crimes against humanity, as recommended in CM/Rec(2022)16.

Third, Twitter does not clarify what measure(s) it will take towards prohibited hate speech as it merely acknowledges that hate speech 'may' be removed.[188] Hence, Twitter's conceptualisation of hate speech and human rights commitments in its ToS do not align with European human rights standards, which require differentiating the elements of criminal hate speech and the removal of criminal hate speech.[189]

Concerning HRDD, similarly to Facebook, Twitter has a dedicated website to explain how it enforces its policies.[190] However, this communication focuses more on the types of enforcement measures rather than explaining the enforcement process (the latter ideally relating to the HRDD process). This means that it is not clear how Twitter approaches its HRDD responsibilities

184 Matsuda (n 32) 6.

185 Twitter Help Center, 'Hateful Conduct' (n 182).

186 Dias Oliva and others warn about the dangers of having automated content moderation tools that disregard pro-social content and reinforce harmful biases. They showed how an AI tool called 'Perspective' developed by Jigsaw considered non-hateful intended slurs used by the drag queen community in the USA more harmful than posts by white nationalists. Thiago Dias Oliva, Dennys Marcelo Antonialli and Alessandra Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' (2021) Sexuality & Culture 25, 700. https://doi.org/10.1007/s12119-020-09790-w.

187 Twitter Help Center, 'Hateful Conduct' (n 182).

188 Ibid.

189 CM/Rec(2022)16), Paragraphs 11 and 32 respectively.

190 Twitter Help Center, 'Our range of enforcement options' (2023) available at <https://help.twitter.com/en/rules-and-policies/enforcement-options> accessed 6 April 2023.

since it does not inform users about the processes in place to 'identify, mitigate and cease' potentially adverse impacts on human rights.

3.5.3 YouTube

YouTube, owned by Google LLC, is a video-sharing social media platform based in the USA, ranked the second most visited website as of October 2022.[191] As of January 2023, YouTube had over 2.5 billion users,[192] with over 400 million active in Europe,[193] and, as of February 2022, a revenue of $28.8 billion.[194] This company qualifies as a VLOP as per the DSA and as a video-sharing platform service as per the AVMSD, and falls under the CSDDD.

YouTube prohibits hate speech, which it defines as 'content promoting violence or hatred against individuals or groups based on any of the following attributes: age; caste; disability, ethnicity; gender identity and expression; nationality; race; immigration status; religion; sex/gender; sexual orientation; victims of a major violent event and their kin; veteran status.'[195] In reviewing YouTube's conceptualisation of hate speech against European human rights standards on countering online hate speech, three comments are in order. First, similar to Facebook and to Twitter, YouTube does not adopt an open-ended list of categories protected from hate speech and therefore fails to align with the open-ended interpretation of protected categories articulated in Article 21 of the CFREU.

Second, YouTube broadly defines hate speech acts as 'encouragement to violence, threats, and incitement to hatred'.[196] Under 'other types of content' YouTube clarifies that dehumanising, alleging superiority or calling for the subjugation or domination over individuals is prohibited. This aligns with Paragraph 11 of CM/Rec(2022)16, which also considers discrimination as one

191 Similarweb (n 157).

192 Statista, 'Most popular social networks worldwide as of January 2023, ranked by number of monthly active users' (2023) available at <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> accessed 6 April 2023.

193 Musically (n 180).

194 Alex Weprin, 'YouTube Ad Revenue Tops $8.6B, Beating Netflix in the Quarter' (*The Hollywood Reporter*, February 2022) available at <https://www.hollywoodreporter.com/business/digital/youtube-ad-revenue-tops-8-6b-beating-netflix-in-the-quarter-1235085391/> accessed 6 April 2023.

195 YouTube, 'Hate speech policy' (2023) available at <https://support.google.com/youtube/answer/2801939?hl=en> accessed 6 April 2023; YouTube, 'How does YouTube protect the community from hate and harassment? (2023) available at <https://www.youtube.com/intl/ALL_ca/howyoutubeworks/our-commitments/standing-up-to-hate/> accessed 6 April 2023; Google, 'Featured Policies: Hate Speech' (2023) available at <https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en> accessed 6 April 2023.

196 YouTube, 'Hate speech policy' (n 195).

of the most serious forms of hate speech. Nevertheless, similar to the Guidelines of Twitter, YouTube too does not include a prohibition of denial or trivialisation of war crimes or crimes against humanity, as recommended in CM/Rec(2022)16. Additionally, YouTube allows hate speech if used for educational purposes.[197] In this context, considering the European human rights standards stemming from the ECtHR decision in *Roj TV A/S v. Denmark*, it should be noted that 'the one-sided coverage with repetitive incitement to participate in fights and actions, incitement to join the organisation/the guerrilla, and the portrayal of deceased guerrilla members as heroes, amounted to propaganda for (…) a terrorist organisation'.[198] Hence, for hate speech to be intended as educational, the authors of such posts must expressly demonstrate that intent and disassociate themselves from such hateful messages. The wrongful implementation of this policy by YouTube was closely scrutinised when in 2021 Syrian activists denouncing air strikes and militant takeovers saw their videos being removed by automated AI content moderation tools.[199] Hence, a clarification aligned with the standards explained in *Roj TV A/S v. Denmark* would contribute to a clearer and more coherent legal framework upholding fundamental rights and protecting human rights activists.[200]

Third, and again similarly to Facebook and Twitter, YouTube too does not clarify what happens to hate speech posted on its platform. YouTube's policies provide incoherent explanations ranging from 'in some rare cases, we may remove content', followed by a requirement of 'repetition' of abusive behaviour,[201] while slightly after this it informs its users that content violating the hate speech policy will be 'removed'. This is an unclear framework and it does not align with CM/Rec(2022)16), which expressly requires platforms to remove criminal hate speech. Nevertheless, it should be noted that YouTube informs users that it may 'limit features' when content comes close to hate speech. This policy aligns with paragraphs 22 and 23 of CM/Rec(2022)16), which require alternative measures (aside from removal) for hate speech that is not criminal, such as deprioritisation or contextualisation.

Like Twitter, YouTube has a dedicated website communicating its enforcement guidelines[202] but it does not address the HRDD process. YouTube is arguably one step behind Twitter, since it seems to place the burden of

197 Ibid.

198 *Roj TV A/S v. Denmark*, App No 24683/14 (ECtHR 17/04/2018), paragraph 46.

199 Kate O'Faherty, 'YouTube keeps deleting evidence of Syrian chemical weapon attacks' (*Wired*, 26 June 2018) available at <https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video> accessed 6 April 2023.

200 TRT World, 'Activists accuse YouTube of destroying digital evidence of Syria war' (2021) available at <https://www.trtworld.com/life/activists-accuse-youtube-of-destroying-digital-evidence-of-syria-war-44809> accessed 6 April 2023; Al Khatib and Kayyali (n 11).

201 YouTube, 'Hate speech policy' (n 195).

202 YouTube Help, 'Reporting and enforcement' (2023) available at <https://support.google.com/youtube/topic/2803138?hl=en&ref_topic=6151248> accessed 6 April 2023.

identifying adverse impacts on human rights on its users rather than on the company itself. This is visible in the 'Reporting' section, which is structured in a way that only reflects options in which the user reports inappropriate content – as opposed to explaining about how YouTube itself proactively and (independently from its users) puts HRDD systems in place to identify adverse impacts on human rights. Table 1 below summarises the findings of the case studies' compliance with the conceptualisation of criminal hate speech.

Table 1 – Case studies' compliance with the conceptualisation of criminal hate speech

Framework	Expressly prohibits HS?	Distinguishes between criminal hate speech and hate speech prohibited by civil or administrative law?	Types of acts in criminal hate speech	Requirement to remove and report criminal hate speech to law enforcement?	Open protected characteristics? Y/N	Protected characteristics	Concerns compared to human rights standards	Positive aspects compared to human rights standards
European human rights standards (essentially from CM/ Rec(2022)16)	Yes	Yes	*- public incitement to commit genocide, violence or discrimination;* *- threats;* *- public insults;* *- public denial, trivialisation and condoning of genocide, crimes against humanity or war crimes;* *- intentional dissemination of HS.*	Yes	Yes: CM/ Rec(2022)16 uses 'such as'.	*Racist, xenophobic, sexist and LGBTI-phobic*, among others.	- No acknowledgment of the intersectionality of systems of historical or systematic oppression.	N/A
Meta/ Facebook	Yes	No. Distinguishes between 3 tiers of severity but consequences of such classification are not clarified in ToS.	*Attacks defined as:* *- violent speech;* *- dehumanising speech;* *- harmful stereotypes;* *- statements of inferiority;* *- expressions of contempt, disgust or dismissal;* *- cursing;* *- calls for exclusion or segregation.*	No	No	*Race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease.* *Age only when used in combination with main characteristics.*	- Allows for gender-based HS during romantic breakups; - Allows for HS towards people who committed violent crimes or sexual offenses (e.g. Rohingya); - Allows for HS if addressed to less than half of a group.	- Recognises that hate speech is used to perpetuate historical oppression.
Twitter	Yes	No	*- Violence;* *- directly attack* *- threat;* *- abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalised.*	No	No	*Race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.*	- Threats need to be 'violent'; - Slurs need to be 'repeated'; - Removal of HS is not certain; it may be required; - Prohibition of denial, trivialisation of war crimes or crimes against humanity not included.	- Recognises that hate speech is used to perpetuate historical oppression.

Framework	Expressly prohibits HS?	Distinguishes between criminal hate speech and hate speech prohibited by civil or administrative law?	Types of acts in criminal hate speech	Requirement to remove and report criminal hate speech to law enforcement?	Open protected characteristics? Y/N	Protected characteristics	Concerns compared to human rights standards	Positive aspects compared to human rights standards
YouTube	Yes	No	*- content promoting violence;* *- content promoting hatred against individuals or groups* *- threats* *- dehumanising, alleging superiority or calling for the subjugation or domination*	No	No	*Age; caste; disability, ethnicity; gender identity and expression; nationality; race; immigration status; religion; sex/gender; sexual orientation; victims of a major violent event and their kin; veteran status.*	- Removal of HS is not certain; it may be removed in rare cases; - Documentary about a hate group is allowed if not promoting hate (possible conflict with Roj TV A/S v. Denmark, if content creators do not expressly disassociate themselves from HS).	N/A

3.6 CONCLUSION

This Chapter addresses a vacuum in the legal framework by clarifying corporate human rights responsibilities in Europe to counter the most serious forms of online hate speech. Following (emerging) standards on HRDD, AI and online content moderation at the international and European level, we claim that there is a legal standard emanating from the HRDD framework in the European context prescribing the responsibility for online platforms, particularly for VLOPs, video-sharing platforms and for platforms under the scope of CSDDD, to align their ToS, with the conceptualisation of the criminal hate speech as explained in the European human rights standards. Further, ToS should explicitly reflect the HRDD responsibilities to prohibit, remove and report criminal hate speech to relevant public authorities. ToS can be considered a human rights 'policy commitment' when they include a clear explanation of the platform's commitments to human rights, including the prohibition of criminal hate speech (Section 3.3). This HRDD measure could also form part of the ongoing preventive HRDD responsibilities to address potentially adverse impacts on human rights.

The limitation of the requirement to harmonise and reflect the conceptualisation of *criminal* hate speech is justified by a growing European human rights understanding of criminal hate speech as reflected in CM/Rec/(2022)16 from which specific HRDD responsibilities can be developed. It is worth remembering that the EC proposed to add hate speech to the list of EU crimes which, if and when this proposal materialises, will strengthen the need for a standardised conceptualisation of criminal hate speech in online platforms' ToS. This legal avenue supports compliance with the transparency and clarity required by Terms and Conditions (Article 14 of the DSA) generally imposed on all providers of intermediary services. To follow an approach of legal coherence, platforms should explicitly conceptualise hate speech in a manner that protects an open-ended list of protected characteristics that have been historically subject to oppression (Section 3.2). This conceptualisation specifically addresses the rights of people or groups of people that have been and remain marginalised members of society.

The three case studies in Section 3.5. demonstrate that although Facebook, Twitter and YouTube have each adopted ToS prohibiting hate speech to a certain degree, none of them currently conceptualises hate speech in a way that is consistent with European human rights standards. More specifically, none recognises the difference between prohibited hate speech and criminal hate speech, nor the specific HRDD responsibilities associated with countering criminal hate speech. Furthermore, the three case studies reveal the lack of alignment of content moderation practices by online platforms with the HRDD responsibilities to identify, mitigate, cease, remedy and inform about potentially adverse impacts on human rights.

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

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