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

Nave, E.V.R.

Countering online hate speech

How to adequately protect fundamental rights?

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 3 juli 2025
klokke uur 16.00 uur

door

Eva Vaz Rodrigues da Nave

geboren te Guarda, Portugal

in 1991

Promotoren:	prof. dr. S. van der Hof prof. dr. T.E. McGonagle
Promotiecommissie:	prof. dr. ir. B.H.M. Custers prof. dr. L.A. van Noorloos dr. H.C. Lahmann prof. dr. drs. J.D. Temperman (Erasmus Universiteit) prof. dr. N. Suzor (Queensland University of Technology)

Opmaak binnenwerk: Anne-Marie Krens – Tekstbeeld – Oegstgeest
Omslagontwerp: Primo!Studio – Delft
Drukwerk: Ipskamp Printing – Amsterdam

ISBN 978 94 6473 832 2

Para a Ana, o Fernando e o Poppy

Acknowledgements

I want to thank my supervisors, Simone van der Hof and Tarlach McGonagle, for sharing their valuable guidance and encouraging me to think critically, key to successfully complete this thesis. Within eLaw, I found inspiring and supporting colleagues. Nikki Vosters, thank you for being my Ph.D. sister, our thesis sessions and chats at Oba created authentic and warm memories. Dimitra Stefoudi, our online writing sessions in the first PhD year, during the pandemic, were life-saving. Kimia Heidary, thanks for contributing to a fun organization of the 2021 eLaw Colloquium and for your advice on the final steps leading up to the defence. Regina Noort and Francien Dechesne, thank you for your kindness and for your check-ins. Gianclaudio Malgieri, your commitment towards the research community at eLaw and your mentorship navigating publication and the research visit in Milan marked exciting times in my Ph.D. journey. Alan Sears, thanks for your patience to introduce me to the layered publishing American system as well as for your advice on conferences and prizes. Henning Lahmann, thanks for your overall support and for taking part of this research to your classes' discussions, that encouraged me to progress and to be more confident in my work at a challenging time. Amara García Adán and Zuzanna Gulczyñska, your visits at eLaw were refreshing and fun times as a more solid Ph.D. cohort.

I am forever grateful to Lottie Lane for accepting to collaborate with me in the second article; your availability, as well as your considerate and constructive feedback were key to decide to continue the Ph.D. journey and, now, I am also lucky to call you a friend. I also want to thank Stephan Raaijmakers and Thijs Veugen for the inspiring discussions during my secondment at the Netherlands Organization for Applied Scientific Research (TNO) and for the fruitful collaboration in this research. I would also like to thank Bart Custers, Marloes van Noorloos, Henning Lahman, Jeroen Temperman, and Nicolas Suzor for their time and thoughtful comments as members of the doctorate committee.

This thesis is also the result of many thought-provoking interactions with networks of other scholars. A huge thank you to my Ph.D. colleagues and friends from the Network for Training on Hate (NETHATE) project; this was a difficult topic, especially for those personally impacted by it, and I will always thank the opportunity to have learnt from you. During these Ph.D. years, I also benefited from the events organized by the Netherlands Network for Human Rights Research (NNHRR), a dynamic community where I met

like-minded scholars across different universities in the Netherlands; this is no doubt an initiative to be praised and I hope it continues to serve future young researchers.

I am grateful for the opportunity to have completed two research visits during my Ph.D. First, I want to thank Marco Pedrazzi and Federica Favuzza for hosting me at the University of Milan La Statale, and Federica Falconi for her availability to discuss this work. Here, I was lucky to be invited by Angelica Bonfanti to attend her course on Business and Human Rights, which significantly influenced the last article of this thesis. During this research visit, I was also very fortunate to be welcomed by Oreste Pollicino and Giovanni De Gregorio, from the Bocconi University, to join local workshops around the topics of public and private values in the digital age where I witnessed senior scholars' moving commitment to guide and collaborate with junior researchers on the academic debate. A thanks to Pietro Dunn and to Francesca Cagossi for being engaging friendly peers in Milan.

Second, I want to thank Katharine Gelber for hosting me at the University of Queensland; this was a memorable experience that I will forever cherish, thank you for your availability and interest to engage with my thesis, for kindly sharing so many literature sources and for connecting me with inspiring scholars. I want to thank Nicholas Carah for the inviting atmosphere at the Center for Digital Cultures, at the University of Queensland, and to thank Axel Bruns and Daniel Angus for the prompt welcome to join events at the Digital Media Research Centre, at the Queensland University of Technology; here, I am also thankful for the stimulating exchanges with Ariadna Matamoros-Fernández. During this research visit, I was lucky to have met Nicolas Suzor; thank you for welcoming me in your group discussions at the ARC Centre of Excellence for Automated Decision-Making and Society and for the provocative literature references. In Brisbane, I was lucky to have Maria Brown and Sara Roetman as lively peer companions.

I want to express my sincere gratitude towards Ana Paiva and her team at the Cabinet of the Portuguese Secretary of State for Science; for almost a year, I was fortunate to count with your encouragement and with the flexibility to accommodate the needed time to complete this thesis. My working experiences prior to this Ph.D. also had a huge influence in successfully completing it. I want to thank extraordinary mentors such as Celine Francois, Camille Aubourg, Aida Ariño-Fernández, Flora Sutherland and Abigail Hartley, you are my best examples of caring and rigorous leadership and your imprint has permeated through this thesis.

I need to thank my therapist, who guided me out of difficult Ph.D. times and helped me choose effective coping strategies.

I would like to thank my friends. Over the past years living in Amsterdam, I am grateful for Nahal's friendship; our chats, work and yoga sessions reinstated a healthy sense of routine key in balancing the research lifestyle. I want to thank the Sabbatical-love crew, Elisa, Kim, Vítor, André and Clélia,

our food and dance gatherings were always heartwarming moments. For over 15 years, my university friends from CBT have remained a fundamental place of joy and playfulness; I hold Guidels, Marissol, Jujineide, Gostini, Tonia, Tomi, Vidalz, Sara, Marinel and Zezi deeply in my heart. I should note that the seed for this thesis was first planted after a conversation with Maria; thank you for keeping my thoughts in check and alive. Other essential friendships have grounded my personality and continue to be one of the love assets that I turn to, and were, naturally, key to finish this Ph.D. Having grown up together in Guarda, I want to thank Inês, Catarina, Rita, Sofia, Ana, Rui, and Yman, our shared teenage years and long lasting friendships into adulthood have been safe havens for laughter and easiness. Among others, I am also so lucky to have in my life Marilda, Inês and Nuno, Shatakshi, Diana and Mega, Tâmara, Leonor, Paradise by proxy, Tomás and Kathryn, Joep and Ararat, Giuliana and Angelo, Daan and Linda, Rose, Filipe, Galamba and Carla, Jane, Lara and Alex, Noffar and Chip, Nana, Chema, Rodrigo and Esteban.

I would like to express my profound gratitude to my family. Tenho toda a sorte do mundo em ter dois incríveis sólidos núcleos familiares, sempre prontos para entrar em acção e para arranjar tempo para uma refeição juntos cheia de entretenimento e partilhas boas durante as visitas a Portugal; amo os Espigados e os Nave. Quero agradecer especialmente à minha irmã Sara e ao meu tio Júlio, que em tempos difíceis me deram especial consolo. Quero agradecer também aos meus avós Júlio, "Manela", Dina, "Zé" António, João e bisavó Patrocínio, que pelo seu árduo trabalho conseguiram ir garantindo melhores condições de vida aos seus e que tanto me ensinaram a crescer. Um obrigada especial ao meu avô "Zé" António, um exemplo raro para a sua geração de patriarca congregador, carinhoso, trabalhador e curioso pela descoberta das novas possibilidades.

Dedico esta tese à minha mãe Ana. Palavras serão sempre insuficientes para te agradecer por me ensinares o amor, a resiliência, o compromisso, o movimento, as plantas, a música. A tese de mestrado não teria sido publicada se não fosse o teu encorajamento e, sem esse passo, dificilmente teria chegado a este doutoramento. És e serás sempre a minha referência. Que sorte ser tua filha!

Esta tese é dedicada também ao amor da minha vida e meu melhor amigo, o Fernando. Desde teres encontrado o anúncio desta posição, passando pela preparação para a entrevista, até todas as questões sobre o que é fazer investigação. *Garantidamente, esta tese – eu própria – não existiria sem ti.* Contigo aprendo a ponderação, o empenho, o humor e a coragem para experimentar o desconhecido. Que saibamos sempre preservar este lugar bonito que temos, no futuro com o nosso filho :)

Table of contents

List of figures and tables

Acronyms

ADI/MSI-DIS	Committee of Experts on combatting hate speech, Council of Europe
AI	Artificial Intelligence
AI Act	Artificial Intelligence Act
APCC	Additional Protocol to the Convention on Cybercrime
AVMSD	Audiovisual Media Services Directive
CFREU	Charter of Fundamental Rights of the EU
CM	Committee of Ministers, Council of Europe
Code or CoC	EU Code of Conduct on countering illegal hate speech online
CoE	Council of Europe
CRT	Critical Race Theory
CSAM	Child Sexual Abuse Material
CSAR	EU Regulation on Child Sexual Abuse Material
CSDDD	Directive on corporate sustainability due diligence
DSA	Digital Services Act
E2EE	End-to-end encryption
EC	European Commission
ECHR	European Convention for the Protection of Human Rights and Fundamental Freedoms
ECJ	European Court of Justice
ECRI	European Commission against Racism and Intolerance
ECtHR	European Court of Human Rights
EDRi	European Digital Rights
EP	European Parliament
EU	European Union
FHE	Fully Homomorphic Encryption
GDPR	General Data Protection Regulation
GIFs	Graphic Interchange Format
GPR	General Policy Recommendation
HRDD	Human Rights Due Diligence
ICC	International Criminal Court
ICCPR	International Covenant on Civil and Political Rights
ICERD	International Convention on the Elimination of All Forms of Racial Discrimination
IMCO	Committee on Internal Market and Consumer Protection
LGBTIQ+	Lesbian, Gay, Bisexual, Trans, Intersex and Queer+
NETHATE	Network of Excellence for Training on Hate
NetzDG	Network Enforcement Law
OECD	Organization for Economic Cooperation and Development

TCOR	EU Regulation on Terrorist Content Online
TEU	Treaty of the European Union
TFEU	Treaty on the Functioning of the European Union
ToS	Terms of Service
UDHR	Universal Declaration of Human Rights
UN	United Nations
UNGPs	United Nations Guiding Principles on Businesses and Human Rights
VLOPs	Very Large Online Platforms
VLOSEs	Very Large Online Search Engines

1 Introduction[1]

This chapter introduces the thesis by presenting the context and social relevance (Section 1.1), the problem statement and research questions (Section 1.2), the methodology (Section 1.3), the scope (Section 1.4), and the outline of the study (Section 1.5).

1.1 Context and social relevance

Human rights activists, civil society, and journalists, all have shared a continued concern with heightened levels of incitement to discrimination and violence towards marginalized communities.[2] These acts of incitement to discrimination and violence towards marginalized groups can be broadly referred to as *hate speech*. Though there is no legally binding definition of hate speech in international or European human rights, *hate speech* has served as an umbrella legal term to describe expressions that "incite, promote, spread, or justify violence, hatred, or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race",[3] colour, language, religion, nationality or ethnic origin, age, disability, sex, gender identity and sexual orientation".[4] In recent years, all across the globe hate speech has risen, for example, towards racialized people such as Black and Asian communities, towards Dalits, Rohin-

1 This research contains content that some readers may find disturbing. Reader discretion is advised. This research was updated in August 2024. Given the fast-developing nature of the regulation in the field of law and digital technologies, readers are advised to confirm legal sources beyond this timeframe. Cross-references should be read as referring to other references within the corresponding Chapter.

2 United Nations Human Rights Council (2021), Report of the Special Rapporteur on minority issues, Report on hate speech, social media and minorities, (A/HRC/46/57), paras 21-28, available at <https://documents.un.org/doc/undoc/gen/g21/054/14/pdf/g2105414.pdf?token=DpvSn1SAikD5BaVA8r&fe=true> accessed 10 April 2024.

3 This thesis rejects theories of different human "races", as all humans belonging to the same species. However, this thesis refers to "race" or "racialized" groups as a means to expose a colonial and imperial process whereby a dominant group ascribes to another group a racial identity for the purpose of continued exclusion and domination.

4 Council of Europe Committee of Ministers (2022), Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech (CM/Rec(2022)16), para. 2.

gya, Roma, and Arabs, and towards religious or belief communities including Baha'i, Muslims, Jews, and Christians.[5]

The rise of hate speech has been documented also in various online environments, including on social media online platforms.[6] These social media online platforms can be broadly described as Internet hosting services that store, moderate, and disseminate user-generated content to the general public through groups, tailored newsfeed, and messaging applications.[7] In other words, online platforms host content generated by users and perform some level of content moderation whereby content is reviewed, promoted, or demoted.[8] With more than half of the world's population as active users of social media online platforms[9] and using these online environments to exercise basic human rights such as freedom of expression, freedom of assembly and association, the content disseminated and moderation policies implemented are increasingly impactful on a global scale.

The proliferation of hate speech on online platforms was initially associated with communication features such as the facilitation to communicate anonymously, the almost instantaneous posting of user-generated content, and the easy dissemination to large audiences.[10] Human rights activists, whistle-

5 A/HRC/46/57 (n 2), para. 24. Gender-related and queerphobic hate speech has also been rising. United Nations Women, FAQs: Trolling, stalking, doxing and other forms of violence against women in the digital age, available at <https://www.unwomen.org/en/what-we-do/ending-violence-against-women/faqs/tech-facilitated-gender-based-violence> accessed 18 November 2024; European Union Agency for Fundamental Rights (2024), Harassment and violence against LGBTIQ people on the rise, available at <https://fra.europa.eu/en/news/2024/harassment-and-violence-against-lgbtiq-people-rise> accessed 18 November 2024.

6 A/HRC/46/57 (n 2), para. 24. See also the Committee of Ministers of the Council of Europe, Recommendation (2024)4 on combating hate crime, para. 56, available at <https://search.coe.int/cm#{%22CoEIdentifier%22:[%220900001680af9736%22],%22sort%22:[%22CoEValidationDate%20Descending%22]}> accessed 27 November 2024.

7 A/HRC/46/57 (n 2), para. 21. This thesis employs "social media online platforms", "social media platforms", and "online platforms" interchangeably. European Union (2022), 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 (DSA), available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2065> accessed 10 April 2024. DSA, Art. 3 (i): "'online platform' means a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service (…)". This thesis notes that online platforms embed different features operating on varied technological settings, including different encryption services.

8 Myers West, S. (2018). Censored, suspended, shadow banned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366-4383.

9 Statistica (2024), Worldwide digital population, available at <https://www.statista.com/statistics/617136/digital-population-worldwide/#:~:text=Worldwide%20digital%20population%202024&text=As%20of%20January%202024%2C%20there,population%2C%20were%20social%20media%20users> accessed 10 April 2024.

10 Statistica (n 9).

blowers former employees of online platforms, and even the United Nations, have warned that business models adopted by some social media companies have not only failed to take down but even amplified online hate speech.[11] Further studies alert to the fact that the intensification of hate speech in digital environments can result in offline hate speech and hate crime.[12] The prevalence of hate speech on online platforms is all the more worrisome given the high user base.[13]

Recent examples of hate speech on social media platforms are found both in market dominant and in more niche online platforms. For example, Meta (previously Facebook) is accused of contributing in 2019 to anti-Muslim riots in Sri Lanka and is currently facing legal actions for playing a crucial role in hosting and promoting commentary inciting to genocide of the Rohingya Muslim community in Myanmar in 2017.[14] In October 2021, the former Facebook employee, whistleblower Frances Haugen, released the "Facebook Papers".[15] This collection of Facebook's internal reports revealed that the company has prioritized economic profit over combating hate speech and other public threats.[16] In particular, these papers show how Facebook's hate speech policies have enabled hate speech content to thrive in countries like Afghanistan, Ethiopia, and India. Other online platforms have also been linked to the spread of hate speech leading to offline violence. This is the case of Gab, a platform which, in 2018, hosted hate speech posted by the shooter before the Pittsburgh synagogue mass shooting. 8kun (previously 8chan) has also been linked to white supremacism, alt-right racism, and hate crimes and, in

11 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 21 Feb 2024; Independent International Fact-Finding Mission on Myanmar, 'Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar' (IIFFMM, Detailed findings), 17 September 2018, A/HRC/39/CRP.2.

12 Judit Bayer & Petra Bard, Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches 38 (July 2020), available at <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/ 655135/IPOLSTU(2020)655135_EN.pdf> accessed 21 Feb 2024.

13 E.g., Statistica, Social media – Statistics and Facts, available at <https://www.statista.com/topics/1164/social-networks/#topicOverview> accessed 21 Feb 2024.

14 The Guardian, Michael Safi (2018) Sri Lanka accuses Facebook over hate speech after deadly riots, available at <https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech> accessed 10 April 204; The New York Times (2018) Facebooks admits it was used to incite violence in Myanmar, available at <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html> accessed 10 April 2024.

15 The Wall Street Journal (2021) The Facebook Papers, available at <https://facebookpapers.com/outlet/wall-street-journal/> accessed 10 April 2024.

16 The Wall Street Journal (n 15).

2019, the site was said to host a post justifying the killing in El Paso targeting members of the Latino migrant community.[17]

In reaction to these events and due to pressure by States, human rights activists and civil society, some online platforms started to self-regulate hate speech,[18] to share data about the hate speech prevalence, and to create oversight boards for appeal procedures on content moderation.[19] However, such self-regulatory efforts are often criticized for not aligning with human rights standards. Some of the main points criticized for not aligning with human rights relate to: i) the conceptualization of hate speech applied by platforms; ii) the mechanisms of enforcement for content moderation policies; iii) the remedies available for users to appeal content moderation decisions.[20]

First, the conceptualizations of hate speech adopted by online platforms can be overbroad or underinclusive when compared with human rights standards.[21] On the one hand, overbroad because in some cases, online platforms take down legal content. For example, in Syria and in Palestine, Meta has taken down and deleted legal content posted by human rights activists.[22] Additionally, the platforms' limited investments in languages other than English for content moderation practices has resulted in the lack of resources

17 Independent, Lizzie Dearden (2018) Gab: Inside the social network where alleged Pittsburgh synagogue shooter posted final message, available at <https://www.independent.co.uk/news/world/americas/pittsburgh-synagogue-shooter-gab-robert-bowers-final-posts-online-comments-a8605721.html> accessed 10 April 2024; The New York Times (2019) Minutes before El Paso killing, hate filled manifesto appears online, available at <https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html?action=click&module=Spotlight&pgtype=Homepage> accessed 10 April 2024.

18 E.g., Meta, Transparency Center "Hate speech", available at <https://transparency.meta.com/en-gb/policies/community-standards/hate-speech/> accessed 11 August 2024; Linked In, Help ""Hateful and derogatory content", available at <https://www.linkedin.com/help/linkedin/answer/a1339812> accessed 11 August 2024; X, Help Center "Hateful conduct", available at <https://help.x.com/en/rules-and-policies/hateful-conduct-policy> accessed 11 August 2024.

19 Kate Klonick, "The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression." *Yale LJ* 129 (2019): 2418.

20 Kate Klonick, "The new governors: The people, rules, and processes governing online speech." *Harv. L. Rev.* 131 (2017): 1598.

21 Podcast series by Katie Pentney "Decoding Hate," Episode 2 "The Hate You Tweet" with Tarlach McGonagle, 10 February 2021, funded by OSCE Representative on Freedom of the Media, #SAIFE project, available at <https://www.decodinghatepod.com/episodes/episode-05-the-anywhere-.w orkout-lhgdz-D0JBu> accessed 10 April 2024.

22 *E.g.*, Human Rights Watch (2023) Meta's Broken Promises, Systematic Censorship of Palestine Content on Instagram and Facebook, available at <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and> accessed 10 April 2024; Business and Human Rights Resource Centre (2021) Syria: New report highlights the complicity of multinational tech companies in the regime's human rights violations, available at <https://www.business-humanrights.org/en/latest-news/syria-new-report-highlights-the-complicity-of-multinational-tech-companies-in-the-regimes-human-rights-violations/> accessed 10 April 2024.

to interpret contexts outside the English speaking countries.[23] On the other hand, the conceptualization of hate speech adopted by online platforms can be underinclusive when it fails to take down illegal content such as hate speech.[24] For instance, Meta was reportedly using a conceptualization of "*protected categories*"[25] which disregarded the standards established by international and European human rights to protect marginalized groups.[26] In the past, Meta's definition led to the removal of a post suggesting that "*all* white people were racist" but authorized a post incentivizing the "killing of *radicalized* Muslims" – justifying that "radicalized Muslims" was a sub-group of protected groups, while "all whites" was more generic and therefore deeming it more critical to protect.[27] This example, together with similar others on different online platforms,[28] showcases some of the human rights concerns around the conceptualizations of hate speech adopted by online platforms.

Second, the mechanisms that online platforms use to enforce their self-regulated policies to counter online hate speech are not communicated in a clear and transparent manner to users. Typically, platforms convey the rules of engagement with their services to users through the terms of service, which can be broadly described as the legal agreements between the service provider

23 MIT Technology Review, Karen Hao (2020) We read the paper that forced Timnit Gebru out of Google. Here's what is says, available at <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/> accessed 10 April 2024.

24 Ariadna Matamoros-Fernández and Johan Farkas (2021) "Racism, hate speech, and social media: A systematic review and critique." *Television & new media* 22.2: 205-224; Anat Ben-David and Ariadna Matamoros Fernández (2016) "Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain". *International Journal of Communication*, *10*, 27; Molly Dragiewicz, Jean Burgess, Ariadna Matamoros-Fernández, Michael Salter, Nicolas P. Suzor, Delanie Woodlock, and Bridget Harris (2018) "Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms". *Feminist Media Studies*, *18*(4), 609-625.

25 "Protected categories" or "protected characteristics" are terms used interchangeably in this thesis as human rights terms employed in the context of the prohibition of discrimination clauses. Typically, these clauses refer to "characteristics" so as to illustrate grounds under which individuals cannot be discriminated upon. E.g., Darina S. Kosinova and Arsenii V. Paliiuk (2021) "Prohibition of Discrimination: Concepts, Features and Obligations of the State according to the Convention for the Protection of Human Rights and Fundamental Freedoms". *L. & Innovative Soc'y*, 99.

26 Paloma Viejo Otero (2022) *Governing hate: Facebook and hate speech*. Diss. Dublin City University; Eugenia Siapera and Paloma Viejo-Otero (2021) "Governing hate: Facebook and digital racism." *Television & New Media* 22.2: 112-130.

27 ProPublica, Julia Angwin and Hannes Grassegger (2027) "Facebook's secret censorship rules protect white men from hate speech but not black children", available at <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms> accessed 10 April 2024.

28 Ariadna Matamoros-Fernández (2017) "Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube." *Information, Communication & Society* 20.6: 930-946.

and the user regulating *inter alia* the prohibited content and behaviour.[29] Although online platforms are increasingly required to communicate in a transparent manner, their terms of service and measures employed to enforce these terms and the data shared regarding such mechanisms is not always sufficient to ensure compliance with human rights standards.[30] Examples of this lack of transparency cover mechanisms for users to report content and to appeal removal decisions.[31]

Moreover, as platforms increasingly provide end-to-end encryption (E2EE) services, new regulatory challenges arise. In particular, while E2EE services enable the safe exercise of freedom of expression for human rights activists persecuted by authoritarian regimes, E2EE can also facilitate illegal activities and organized crime, such as hate speech and incitement to violence.[32] While E2EE started by being used on messaging applications such as Signal,[33] WhatsApp,[34] Telegram,[35] E2EE is increasingly provided within messaging features of mainstream online platforms. For example, E2EE is now provided on Meta's Messenger (formerly Facebook)[36] and on X (formerly Twitter).[37]

29 Sandra Braman and Stephanie Roberts (2003) "Advantage ISP: Terms of service as media law." *New media & society* 5.3: 422-448.

30 European Commission, Monitoring rounds of the Code of conduct on countering illegal hate speech online, available at <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 10 April 2024.

31 Aina Turillazzi, Mariarosaria Taddeo, Luciano Floridi, and Federico Casolari. (2023) "The digital services act: an analysis of its ethical, legal, and social implications." *Law, Innovation and Technology* 15.1: 83-106; Ilaria Buri and Joris van Hoboken (2021) "The Digital Services Act (DSA) proposal: a critical overview." Digital Services Act (DSA) Observatory; Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. "What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation." *International Journal of Communication* 13 (2019): 18.

32 Sarvesh Methi Medianama (2022) "How end-to-end encryption impact human rights? The Good and the Bad", available at <https://www.medianama.com/2022/04/223-end-to-end-encryption-human-rights-impact/> accessed 10 April 2024; Business for Social Responsibility (2022) Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption, available at <https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf> and available at <https://www.bsr.org/en/reports/metas-expansion-end-to-end-encryption> accessed 10 April 2024.

33 Signal Support, available at <https://support.signal.org/hc/en-us/articles/360007320391-Is-it-private-Can-I-trust-it> accessed 10 April 2024.

34 WhatsApp Help Center, About end-to-end encryption, available at <https://faq.whatsapp.com/820124435853543?locale=en_US&cms_id=820124435853543&draft=false> accessed 10 April 2024.

35 Telegram FAQ, available at <https://telegram.org/faq#secret-chats> accessed 10 April 2024.

36 Meta (2023) Messenger End-to-End Encryption Overview, available at <https://engineering.fb.com/wp-content/uploads/2023/12/MessengerEnd-to-EndEncryptionOverview_12-6-2023.pdf> accessed 10 April 2024.

37 Lance Whitney (2023) "Twitter rolls out encryption for direct messages but with key limitations", available at <https://www.zdnet.com/article/twitter-rolls-out-encryption-for-direct-messages-but-with-key-limitations/> accessed 10 April 2024.

These contexts impose new debates as to what are the responsibilities of the platforms to remove hate speech from these contexts.

Third, there is a significant lack of clarity as to the liability regimes and remedial responsibilities of online platforms for cases in which their content management decisions caused or contributed to hate speech.[38] As cases of hate speech amplified by online platforms rise and noting that the services provided by these platforms span across multiple jurisdictions, it becomes challenging for individuals wanting to bring claims against these companies to identify the judicial system competent to judge such cases.[39]

Aside from the debates about the competent jurisdictions, platforms have also refused to provide the remedies requested by people targeted by hate speech amplified by their services.[40] For example, reports by the United Nations and by Amnesty International show that the content management algorithms (including content moderation, ranking, and recommendation algorithms) utilized by Meta amplified criminal hate speech in the form of incitement to genocide towards the Muslim Rohingya community in Myanmar.[41] Meta, though acknowledging to a certain degree that its content management contributed to amplifying violence in Myanmar, does not recognize its remedial responsibilities towards the survivors of the Rohingya community.[42] The case of Myanmar is one of the best documented ones, but there is evidence indicating the need to study similar contributions of online platforms to the proliferation of criminal hate speech in different contexts (*e.g.* the amplification of hate speech on online platforms towards the Roma community in Europe, towards Baha'i, Muslim, Christian, Dalits, migrants, and racialized communities in many countries.[43] The platforms' refusal to remediate the harms caused to these communities illustrates the need for clearer regulation of the remedial responsibilities of online platforms in the context of criminal hate speech.

Moreover, studies show that platforms have prioritized user engagement over human rights resulting in online hate speech spreading much faster, farther, and reaching a much wider audience than innocuous content on online

38 Amnesty International (2022) 'Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya', available at <https://www.amnesty.org/en/documents/asa16/5933/2022/en/> accessed 10 April 2024.

39 Julia Hörnle (2021). *Internet jurisdiction law and practice*. Oxford University Press.

40 Justice Belen Galvan (2020) "Facebook's Legal Responsibility for the Rohingya Genocide". *USFL Rev.*, *55*, 123.

41 Amnesty International (n 38).

42 Amnesty International (n 38), 55 and Chapter 9; Neema Hakim (2020) "How social media companies could be complicit in incitement to genocide." *Chi. J. Int'l L.* 21: 83; Kyle Rapp (2021) "Social media and genocide: The case for home state responsibility." *Journal of Human Rights* 20.4: 486-502.

43 A/HRC/46/57 (n 2), paras. 35-40.

platforms.[44] This context poses additional conceptual questions related to whether platforms should be required to adjust their business models to prioritize human rights instead of user engagement and profit.

In an effort to regulate and to democratically oversee the regulatory frameworks established by businesses to counter online hate speech, both States and international and regional organizations have been producing sector-specific legal and standard-setting instruments. Some instruments deal with the conceptualization of hate speech[45] and others with the conceptualization of the corporate human rights responsibilities strengthening online platforms' responsibilities to counter online hate speech.[46] Additionally, legislators have worked with online platforms to develop co-regulatory approaches to countering online hate speech.[47] Nevertheless, discussions have arisen regarding the effectiveness and adequacy of such regulatory frameworks in promoting the respect for the human rights of people targeted by hate speech.[48]

For example, Germany adopted in 2017 the Network Enforcement Law (NetzDG) which requires companies to remove "manifestly unlawful" content

44 Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee (2019) "Spread of Hate Speech in Online Media", *Proceedings 10th ACM Conf. on Web Science*; United Nations Report of the Special Rapporteur on minority issues (2021) A/HRC/46/57.

45 E.g. Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950; European Union, Charter of Fundamental Rights of the European Union, 2012/C 326/02, 26 October 2012 (CFREU); Council of Europe, Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech (CM/Rec(2022)16); 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 (EU Framework Decision), 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 (AVMSD).

46 *E.g.* 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 (DSA), AVMSD, 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, AI Act), European Union, 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 (CSDDD).

47 European Commission, Code of Conduct to counter illegal hate speech online, 2016 available at<https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 28 August 2024, [hereinafter EU Code of conduct].

48 *E.g.* Natalie Alkiviadou (2019) "Hate speech on social media networks: towards a regulatory framework?." Information & Communications Technology Law 28.1: 19-35.

in 24 hours, once reported by users.[49] Failing to do so can result in fines up to 50 million euros.[50] Although a pioneer legislation in the fight against hate speech, this law promotes the removal of content without due consideration, and critics claim that it enables unnecessary and disproportionate restrictions of the right to freedom of expression.[51] Additionally, requiring users to report content lacks the understanding that the online world is a polarized one: hate speech occurring in so-called *echo chambers*[52] and *filter bubbles*[53] will not likely be reported as online networks are encouraged to gather like-minded people. In another example, the Parliament of the United Kingdom adopted in 2023 the Online Safety Act which creates a duty of care for online platforms towards their users.[54] This Act has been severely criticized for, among others, enabling online platforms to take down both illegal and "lawful but harmful" content.[55] To be more specific, in the case of due diligence duties to prevent children from being exposed to harmful content, one of the possible measures is content

49 Germany, "Act to Improve Enforcement of the Law in Social Networks" (Network Enforcement Act), Ministry of Justice and Consumer Protection, 12 of July 2017, English Version, available at: <https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/RefE/NetzDG_engl.pdf?__blob=publicationFile&> accessed 12 January 2025. Heidi Tworek and Paddy Leerssen. "An analysis of Germany's NetzDG law." First session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (2019).

50 Heidi Tworek and Paddy Leerssen (n 49).

51 *E.g.* Diana Lee (2017) "Germany's NetzDG and the threat to online free speech." Yale Media Freedom & Information Access Clinic Case Disclosed Blog 10, available at <https://law.yale.edu/mfia/case-disclosed/germanys-netzdg-and-threat-online-free-speech> accessed 10 April 2024; Jillian York (2022) *Silicon values: The future of free speech under surveillance capitalism*. Verso Books; Shoshana Zuboff (2019) "Surveillance capitalism and the challenge of collective action." *New labor forum*. Vol. 28. No. 1. Sage CA: Los Angeles, CA: SAGE Publications; Susie Alegre (2022) *Freedom to Think: The Long Struggle to Liberate Our Minds*. Atlantic Books; Frank Pasquale (2015) *The black box society: The secret algorithms that control money and information*. Harvard University Press; David Kaye (2019) "Speech police: The global struggle to govern the Internet".

52 Terren Ludovic and Rosa Borge-Bravo Rosa Borge-Bravo (2021) "Echo chambers on social media: A systematic review of the literature." *Review of Communication Research* 9; Walter Quattrociocchi, Antonio Scala, and Cass R. Sunstein (2016) "Echo chambers on Facebook." *Available at SSRN 2795110*.

53 Dominic Spohr (2017) "Fake news and ideological polarization: Filter bubbles and selective exposure on social media." *Business information review* 34.3: 150-160; Uthsav Chitra and Christopher Musco (2020) "Analyzing the impact of filter bubbles on social network polarization." *Proceedings of the 13th International Conference on Web Search and Data Mining*.

54 Markus Trengove et al. (2022) "A critical review of the Online Safety Bill." *Patterns* 3.8. Alexander Dittel (2022) "The UK's Online Safety Bill: The day we took a stand against serious online harms or the day we lost our freedoms to platforms and the state?." *Journal of Data Protection & Privacy* 5.2: 183-194.

55 Tech Against Terrorism's submission to the United Kingdom Draft Online Safety Bill Consultation (2021), available at <https://www.techagainstterrorism.org/wp-content/uploads/2021/09/Tech-Against-Terrorisms-Response-%E2%80%93-Joint-Committee-OSB-call-for-written-evidence.pdf> accessed 10 April 2024.

removal.[56] Critical views emphasize that such a framework undermines the rule of law and due process and has the potential to legitimize the censorship of legal content.[57] These examples of regulation passed by States[58] illustrate challenges to regulate the responsibilities of online platforms to counter online hate speech in a human rights compliant approach, namely in compliance with the right to freedom of expression, prohibition of discrimination, right to a fair trial, and to an effective remedy.[59]

At the level of the European Union (EU), in 2022, the European Parliament and the Council of the EU adopted the Digital Services Act (DSA).[60] The DSA aims to establish a harmonized approach within the EU Member States to counter illegal content online, clarify the transparency responsibilities of online platforms, and counter disinformation.[61] Some of the key transparency responsibilities include making publicly available annual reports communicating the number of orders that they received from national authorities, details of their content moderation measures, number of content removed, and the accuracy and error rate of content moderation automated systems.[62] Before that, in 2016, the European Commission and some of the biggest online platforms had already agreed on a co-regulatory approach to counter online hate speech, i.e., the Code of conduct on countering illegal hate speech online.[63] However, both the DSA and Code of conduct lack clarity on key aspects central to coun-

56 UK Online Safety Act (2023) Provision 12(8) (e) on Safety duties protecting children, available at <https://www.legislation.gov.uk/ukpga/2023/50/section/12/enacted> accessed 11 August 2024.

57 EDRi (2023) Online Safety Bill insecure: international organisations, academics and cyber experts urge UK government to protect encrypted messaging, available at <https://edri.org/our-work/online-safety-bill-insecure-international-organisations-academics-and-cyber-experts-urge-uk-government-to-protect-encrypted-messaging/> accessed 10 April 2024.

58 Other examples of constrains to design domestic regulation on this domain can be found, e.g., in Ireland and France. The Irish government proposed, in December 2020, the Online Safety and Media Regulation Bill to respond to harmful materials online. The expert reactions to this draft Bill have also pointed out to the vagueness of what is to be considered as harmful content and how this lack of clarity could be infringing constitutional rights. See UCD Centre for Digital Policy, available at <https://digitalpolicy.ie/explainer-ireland-sonline-safety-and-media-regulation-bill/> accessed 18 November 2021. France, in May 2021, had adopted the "Avia law" with the purpose to counter online hate speech. However, in June 2021, the Constitutional Council declared the main provisions of this law unconstitutional for infringing freedom of speech and for lacking due regard to the necessity and proportionality legal criteria in restrictions of freedom of expression. E.g., EDRi, "French Avia law declared unconstitutional: what does this teach us at EU level?", available at <https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/> accessed 18 November 2021.

59 For a legal framing within the European context, see ECHR, Articles 6, 10, 13, 14, and 17.

60 DSA (n 7). Martin Husovec and Irene Roche Laguna (2023) "Digital services act: A short primer," *Principles of the Digital Services Act* (Oxford University Press).

61 E.g. DSA, Recital 9.

62 DSA, Art. 15.

63 EU Code of conduct (n 47).

tering online hate speech in a human rights compliant manner. For example, neither the DSA nor the Code of conduct explain the conceptualization of hate speech. Furthermore, both instruments adopt a vague approach to the corporate human rights responsibilities of online platforms to counter online hate speech beyond the transparency requirements, e.g. lacking a clear legal framing of remedial responsibilities.[64] It is against this background, highlighting the rise of hate speech and the complexities of designing human rights frameworks to counter online hate speech, that a problem statement and research questions have been formulated (next section) to further the human rights compliant fight against hate speech.[65]

1.2 Problem statement and research questions

– *Problem statement*

As explained in the previous section, the relationship between the right to freedom of expression and its restrictions in cases of hate speech in the digital environment requires further legal interpretation.[66] The overall aim of this thesis is to critically analyze how the corporate human rights responsibility framework applies to online platforms countering online hate speech in the European context. Hence, the Problem Statement is as follows:

> *Building on a critical conceptualization of online hate speech, and more specifically on criminal hate speech, deriving from the European regulatory and policy framework, how can European legislators, both at the European Union and at the Council of Europe levels, clarify the responsibilities of online platforms to counter online hate speech whilst upholding fundamental rights?*

64 Article 19 (2021) At a glance: Does the EU Digital Services Act protect freedom of expression, available at <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 10 April 2024.

65 This thesis was funded by the Network of Excellence for Training on Hate (NETHATE). Overall, the NETHATE project aimed to overall investigate the societal impact of hate by researching (a) the nature of hate, (b) the dynamics of its spread in both offline and online for a, (c) the impact on victims, and (d) mitigation and reconciliation strategies. Within the NETHATE project, the research presented in this thesis fits under Work Package No. 2 on Technology and Social Media, which specifically focuses on researching the impacts and potential mitigation and reconciliation strategies to online hate speech through technology and social media.

66 Oreste Pollicino & Gabriella Romeo (Eds.). (2016). *The internet and constitutional law: the protection of fundamental rights and constitutional adjudication in Europe*. Routledge, Taylor & Francis Group; Oreste Pollicino (2021). *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (1st ed.). Hart Publishing, available at <https://doi.org/10.5040/9781509912728> accessed 28 August 2024.

In line with the Problem Statement, this study aims to answer four Research Questions, each contributing to addressing the problem statement as set out in the next paragraphs.

– *Research Question 1*

> *To what extent do the main elements of hate speech under European human rights standards align with the conceptualization of hate speech promoted by critical legal theory and do those elements require further clarification?*

By answering Research Question 1, Chapter 2 critically reviews the conceptualization of hate speech developed in European human rights standards and establishes a working definition of hate speech. Within this working definition, this chapter similarly clarifies the working definition in European human rights standards for criminal hate speech, which is the legal basis for the following Research Questions. From the second to the fourth Research Questions, this thesis analyses the European regulatory and policy framework on the corporate human rights responsibilities of online platforms to counter criminal hate speech, and advances legal approaches to clarify and strengthen this framework.

– *Research Question 2*

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

By answering Research Question 2, Chapter 3 clarifies and advances new regulatory approaches to strengthen the human rights responsibilities of online platforms to prevent criminal online hate speech, specifically through the design of terms of service (ToS).[67]

– *Research Question 3*

> *To what extent can an innovative legal interpretation of technological developments clarify and expand the human rights due diligence (HRDD) of online platforms providing end-to-end encrypted (E2EE) services in the European context to not host criminal hate speech in the form of incitement to violence, and to what*

67 "Terms of service", "community guidelines", "terms and conditions" are used interchangeably in this thesis.

extent can such interpretation result in new HRDD responsibility standards for E2EE services' cooperation with law enforcement?

By answering Research Question 3, Chapter 4 clarifies and advances new regulatory approaches to strengthen the human rights responsibilities of online platforms to mitigate criminal hate speech in the specific case of end-to-end encrypted (E2EE) services.

– *Research Question 4*

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

By answering Research Question 4, Chapter 5 clarifies and advances new regulatory approaches to strengthen the human rights remedial responsibilities of online platforms which caused or contributed to criminal online hate speech.

1.3 METHODOLOGY

The overall aim of this thesis is to analyse the European regulatory and policy framework on the corporate human rights responsibilities of online platforms to counter online hate speech, and to advance legal approaches to clarify and strengthen this framework. Three methodologies were employed to answer these Research Questions: doctrinal legal research (Section 1.3.1), comparative legal research (Section 1.3.2), and interdisciplinary legal research (Section 1.3.3).[68] With all methodologies employed across all Chapters, the following paragraphs give an overview of these methodologies and how they were applied throughout the thesis.

1.3.1 Doctrinal legal research

This thesis employs doctrinal research thorough all Chapters.[69] This methodology seeks to systematically interpret, to identify legal loopholes, and to propose analytical approaches to improve the legal coherence of the applicable

68 This thesis takes a methodological approach similar to that of Sabine Witting Ph.D. thesis. See Witting, S. K. (Sabine K. (2020). *Child sexual abuse in the digital era?: Rethinking legal frameworks and transnational law enforcement collaboration.*

69 Mike McConville (2007) Research Methods for Law, Edinburgh; Matyas Bodig (2015) Legal Doctrinal Scholarship and Interdisciplinary Engagement, Erasmus Law Review, Vol. 8.

normative frameworks.[70] The normative framework is composed of binding legal sources and non-binding but authoritative sources. Binding sources encompass regulatory[71] texts and case law. Non-binding legal sources encompass, for example, commentaries to binding legal sources, *travaux préparatoires*, standard-setting policy instruments, and scholarly writing.

This thesis critically examines both regulatory and standard-setting instruments and the relationship between them. The main regulatory and policy frameworks analyzed in this thesis are human rights, corporate human rights responsibilities, and platform governance. These frameworks were selected to ensure coverage of the individual human rights (human rights framework), the States' obligation[72] to regulate and to remediate harms caused by the private sector (platform governance), and the online platforms' responsibilities to respect human rights (corporate human rights responsibilities).

The sources examined through doctrinal legal research in this thesis cover both European regional and international regulatory and policy sources. The regional sources focus on instruments from the Council of Europe and the European Union level, due to the interdependence of these two regional human rights systems. The international sources focus primarily on instruments adopted at the United Nations and the Organization for Economic Cooperation and Development (OECD), because, in the case of human rights due diligence, the European frameworks have been grounded on those international sources.[73] An example of the influence of the international standards on European frameworks is the case of the European corporate human rights responsibilities framework. All sources analyzed in this thesis are publicly available for example on international law databases, judicial case law data bases, scholarly writings, the online platforms' websites, and on performance reports submitted to the European Commission on the implementation of the Code of Conduct on countering illegal hate speech online. No personal data was collected or studied for the purpose of this thesis.

The binding regulatory international legal instruments studied cover mostly the European regional legal framework stemming from the Council of Europe and include: the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR);[74] jurisprudence of the European Court of Human Rights (ECtHR); the Additional Protocol to the Convention on Cybercrime, concerning the criminalization of acts of a racist and xenophobic

70 Mike McConville (n 69) and Matyas Bodig (n 69).
71 "Regulatory" and "statutory" are used interchangeably.
72 The term "obligation" is employed to refer to legally binding standards, whilst the term "responsibility" is employed to refer to non-legally binding standards.
73 See Chapter 3, Section 3.3. on the "Broader framework: AI and the corporate responsibility to protect human rights".
74 ECHR (n 45).

nature committed through computer systems (AP to the Budapest Convention);[75] the Convention on preventing and combating violence against women and domestic violence (the Istanbul Convention).[76] The binding European instruments examined include: the Charter of Fundamental Rights of the European Union (CFREU); jurisprudence of the Court of Justice of the European Union (CJEU); the Directive on corporate sustainability due diligence (CSDDD);[77] the Artificial Intelligence Act (AI Act);[78] the Digital Services Act (DSA);[79] the 2018-revised AVMSD;[80] and, the Council Framework Decision on combating certain forms and expression of racism and xenophobia by means of criminal law.[81]

The policy instruments studied in this thesis cover international sources, with a specific focus on instruments adopted at the European level. The European focus covers instruments both at the Council of Europe and European Union levels. The international policy instruments reviewed include: the United Nations Guiding Principles on Businesses and Human Rights (UNGPs);[82] the OECD Declaration and Guidelines for Multinational Enterprises;[83] the OECD Due Diligence Guidance for Responsible Business Conduct.[84] The Council of Europe policy instruments include: General Policy Recommendations Numbers 7, 11, and 15 by the European Commission against Racism and Intolerance (ECRI);[85] Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech;[86] Recommendation Rec(97)20 of the Committee of Ministers to member States on "hate speech";[87] Recommendation Rec(97)21 of the Committee of Ministers to mem-

75 Additional Protocol to the Convention on Cybercrime, Jan. 28, 2003, E.T.S. 189, available at <https://rm.coe.int/168008160f> accessed 10 April 2024.

76 Convention on Preventing and Combating Violence Against Women and Domestic Violence, May 11, 2011, E.T.S. 210, available at <https://rm. coe.int/1680 08482e> accessed 10 April 2024.

77 CSDDD (n 46).

78 AI Act (n 46).

79 DSA (n 46).

80 AVMSD (n 45).

81 EU Framework Decision (n 45).

82 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.

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

84 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.

85 European Commission against Racism and Intolerance (ECRI) Standards, available at <https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/ecri-standards> accessed 11 April 2024.

86 CM/Rec(2022)16 (n 45).

87 Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech" adopted on 30 October 1997.

ber States on the media and the promotion of a culture of tolerance;[88] and, Council of Europe Explanatory Memorandums.[89] The European Union policy instruments include: the Code of conduct on countering illegal hate speech online;[90] the Directive of the European Parliament and of the Council on combating violence against women and domestic violence;[91] the European Parliament's Resolution on extending the list of areas of crime at the EU level to include hate speech and hate crime and also urging the Council to follow through with the Commission's proposal;[92] reports by the European Commission from the monitoring rounds on the online platforms' compliance with the Code of Conduct on countering illegal hate speech online.

Additionally, this thesis analyses scholarly writing including: legal scholarship (*e.g.*, human rights, platform governance, critical race legal theory, and feminist legal methods[93]); sociological scholarship (*e.g.*, black feminist theory and intersectionality theory); and, computer science and artificial intelligence scholarship (*e.g.*, content moderation algorithms, recommender algorithms, ranking algorithms, encryption systems). These scholarly writings are foundational to better interpret regulatory, policy, and technological mechanisms to countering hate speech on online platforms. Finally, this thesis also examined non-academic publications from international organizations, civil society organizations, and human rights think tanks, which have completed the analysis of regulatory and policy frameworks.

1.3.2 Comparative legal research

Comparative research aims to provide a comprehensive analysis of different normative approaches to the same normative challenge. This methodology seeks to offer a solid understanding of how different legislative systems impact

88 Recommendation No. R (97) 21 of the Committee of Ministers to member states on the media and the promotion of a culture of tolerance adopted on 30 October 1997.

89 Council of Europe, Committee of Ministers (2014) Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a guide to human rights for internet users – Explanatory Memorandum, available at <https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=09000016805c6f85#P118_25200> accessed 11 April 2024. It should be noted that, while not having the same policy impact as recommendations, explanatory memoranda assist in interpreting the recommendations.

90 EU Code of conduct (n 47).

91 European Union, Directive (EU) 2024/1385 of the European Parliament and the Council of 14 May 2024 on combating violence against women and domestic violence, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401385> accessed 20 November 2024.

92 European Parliament (2024) Resolution of 18 January 2024 on extending the list of EU crimes to hate speech and hate crime (2023/2068(INI)) available at <https://www.europarl.europa.eu/doceo/document/TA-9-2024-0044_EN.html> accessed 11 April 2024.

93 Katharine T Bartlett (2018) "Feminist Legal Methods [1990]." Feminist legal theory. Routledge, 370-403.

the normative development and to compare which approach best delivers on the given objective.[94]

This thesis applies, across all Chapters, comparative research in three domains. First, comparative research is employed to clarify and investigate the relationship between the international corporate human rights standards with the European approaches to corporate human rights standards. Second, comparative research is used to clarify and investigate the relationship between the two different European human rights systems, *i.e.* the Council of Europe and the European Union systems. Third, the comparative approach is utilized to clarify and investigate the relationship between the European human rights standards with the standards adopted by the online platforms.

With the view to explore the relationship between the European human rights standards and the standards adopted by online platforms, Chapter 3 employs empirical qualitative legal research. Empirical qualitative legal research aims to advance research findings based on the analysis of the operationalization of law in practice.[95] Chapter 3 employs a set of qualitative criteria for the systematic analysis of the publicly available terms of service adopted by online platforms to evaluate the alignment or lack thereof between these self-regulatory practices with the European human rights standards. More specifically, this thesis examines whether the terms of service of online platforms align with not just the human rights conceptualization of hate speech as well as with the corporate human rights responsibilities framework.

1.3.3 Interdisciplinary legal research

Interdisciplinary legal research aims to connect the research findings from the legal scholarship with non-legal disciplines.[96] This methodology stems from the recognition of the added value of reviewing legal research through the lens of the questions, methods, and outcomes from other disciplines.[97] This thesis employs interdisciplinary legal research by combining legal research with sociology and computer science studies.

Building on sociology studies, this thesis explores black feminist theory and intersectionality theory. Chapter 2 critically analyses regulatory and policy

94 Mark Van Hoecke (2015) Methodology of Comparative Legal Research, Law and Method, available at<https://www.bjutijdschriften.nl/tijdschrift/lawandmethod/2015/12/RENM-D-14-00001.pdf> accessed 18 November 2024.

95 Edward L Rubin (1997) "Law and the Methodology of Law." Wis. L. Rev.: 521; Aikaterini Argyrou (2017) "Making the case for case studies in empirical legal research." *Utrecht Law Review* 13.3: 95-113. Lisa Webley (2010) "Qualitative approaches to empirical legal research."

96 Andria Naudé Fourie (2015) "Expounding the place of legal doctrinal methods in legal-interdisciplinary research." Erasmus L. Rev. 8: 95.

97 Robert C Clark (1981) "The interdisciplinary study of legal evolution." *The Yale Law Journal* 90.5: 1238-1274.

frameworks in light of the conceptualization of critical race theory and black feminist intersectionality theory. These two legal and sociology theoretical frameworks were selected because they gave prominence to the term (racist) "hate speech" from the perspective of the people targeted by it, which grounded the critical view adopted in this thesis.

Building on computer science studies, Chapter 4 investigates the technical and legal frameworks applicable to content moderation algorithms, recommender algorithms, ranking algorithms, and encryption systems. In particular, Chapter 4 examines the principles guiding the development and deployment of algorithms used by online platforms to manage content with the goal of developing a more practice-based regulatory approach.

1.4 SCOPE

The scope of analysis in this thesis focuses primarily on the European regional regulatory and policy instruments. The main reason for the adoption of a European approach stems from the growing development in the European context of both regulatory and policy initiatives to govern the responsibilities of online platforms to counter online hate speech. This fast development of regulatory and policy frameworks presents a pressing need to clarify the interplay between the various instruments as well as to clarify means of addressing existing legal incoherences and loopholes. This European regional approach considers instruments at both the Council of Europe and the European Union level. There are two main reasons for the combined consideration of these two European regimes. Firstly, there is an overall alignment of key human rights values between the Council of Europe and the EU. For example, as per Art. 52 (3) of the CFREU, provisions in the CFREU should be interpreted with equal meaning to corresponding provisions in the ECHR.[98] Secondly, noting that the EU is due to accede to the ECHR[99] and that negotiations have resumed in 2020,[100] the EU human rights system will be bound to follow the human rights framework of the Council of Europe. To complement the European focus, this thesis occasionally considers the international human rights standards, whenever guidance is lacking or unclear at the European level. For example, given that the UNGPs find no European instrument with equivalent scope, this thesis draws significantly from it regarding the overall corporate human rights responsibilities framework.

98 CFREU, Art. 52 (3).

99 European Union, *Treaty on European Union (Consolidated Version), Treaty of Maastricht*, Official Journal of the European Communities C 325/5; 24 December 2002, 7 February 1992 (TEU), Art. 6 (2).

100 Council of Europe, European Union accession to the European Convention on Human Rights, available at <https://www.coe.int/en/web/portal/eu-accession-echr-questions-and-answers> accessed 11 April 2024.

The scope of analysis in this thesis considers both regulatory and policy instruments. The main justification for this decision originates from the fact that the domains under analysis are, in some cases, not (yet) accompanied by legally binding regulation. To recall, the three main domains under study are: (1) human rights of individuals targeted by hate speech; (2) the corporate human rights responsibilities of online platforms to counter online hate speech; and, (3) the States' obligation to regulate the human rights, which despite having legal grounding in various international and regional legally binding sources, provides no legally binding definition of hate speech.

In the first domain, this thesis analyses interpretative sources such as standard-setting instruments which help to clarify the meaning of human rights legally binding provisions impacting the regulation of hate speech. The second and third domains, *i.e.* responsibilities of online platforms to counter online hate speech and of States to regulate the platforms responsibilities, have just recently started to be addressed through regulatory approaches. Though these regulatory developments have been taking place mostly at the European regional level, this work also reviews these European regulatory developments in light of the general corporate human rights responsibility framework advanced internationally with the UNGPs and subsequent guidance by the OECD. Figure 1 below provides an overview of the regulatory and policy human rights frameworks impacting the regulation of online hate speech analyses in this thesis.

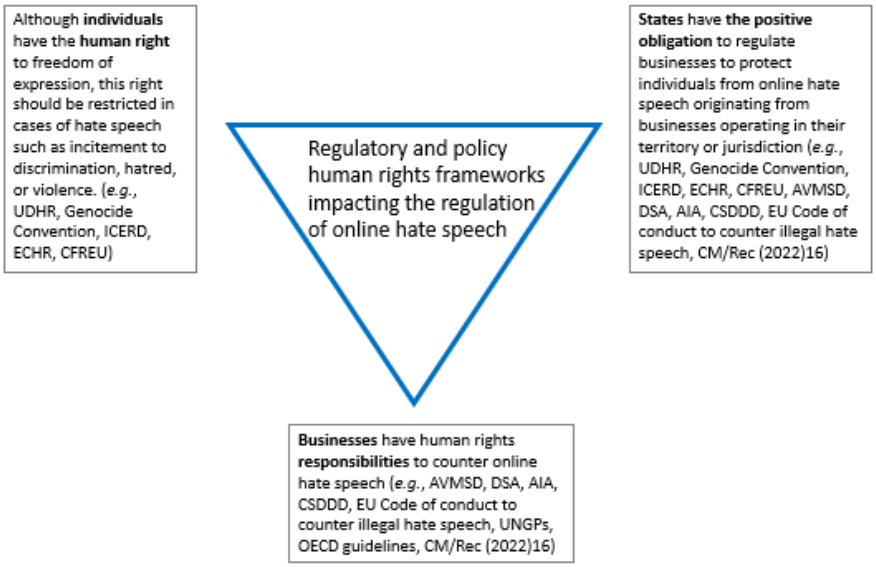

Figure 1 – Overview of regulatory and policy human rights frameworks impacting the regulation of online hate speech.

Another essential clarification regarding the scope of this thesis is that this research focuses exclusively on criminal hate speech for the analysis of the human rights responsibilities of online platforms (Chapters II to IV). Hate speech can be broadly subdivided into two categories *i.e.*, hate speech that is criminally actionable, and hate speech prohibited under civil or administrative law.[101] While hate speech which should be prohibited under civil or administrative law requires contextual interpretation, hate speech that is criminally actionable has a clearer legal and more foreseeable framing.[102] More specifically, the human rights conceptualization of criminal hate speech followed in this thesis derives[103] from established treaty obligations which gather heightened international consensus. This legal clarity and foreseeability framing criminal hate speech lay the foundation for the extrapolation of clearer human rights responsibilities for online platforms.

The final remark regarding the scope relates to the conceptualization of the services provided by online platforms. To recapitulate, online platforms are broadly conceptualized as Internet hosting services that host, moderate, and disseminate content generated by its users to the general public.[104] This work builds on the acknowledgement of three broad types of services provided by online platforms, *i.e.*, hosting, moderation, and dissemination.

This thesis employs the expression "content management" policies to refer to all services provided by online platforms.[105] Importantly, this work deviates from the use of content moderation as an all-encompassing term to refer to the services provided by online platforms.[106] The main reason for this terminology deviation stems from the fact that content moderation represents not all, but one type of service provided by these platforms as it relates to the decision of what content remains or is taken down from a platform. However, given the large user base, the significant amounts of user-generated content, and the platforms' goal of increasing user engagement, online platforms started

101 CM/Rec(2022)16, para. 1 (3) (a) (i) and (ii). This categorisation of hate speech adopted in CM/Rec(2022)16 aligns with the United Nations (2013) Annual report of the United Nations High Commission for Human Rights, A/HRC/22/17/Add.4 (Rabat Plan of Action) available at <https://www.ohchr.org/en/documents/outcome-documents/rabat-plan-action> accessed 11 April 2024, para. 12.

102 European Court of Human Rights (updated on 2022) Guide on Article 7 of the European Convention on Human Rights, No punishment without law: the principle that only the law can define a crime and prescribe a penalty, available at <https://www.echr.coe.int/documents/d/echr/Guide_Art_7_ENG> accessed 11 April 2024, 12-15.

103 Indirectly, given the heavy reliance on CM/Rec(2022)16.

104 DSA, Art. 3 (i).

105 This approach is inspired by the work of Tarleton Gillespie on algorithms employed by social media online platforms beyond content moderation. See *e.g.* Tarleton Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

106 Nicolas Suzor (2020) "Understanding content moderation systems: new methods to understand internet governance at scale, over time, and across platforms." In *Computational Legal Studies,* pp. 166-189. Edward Elgar Publishing.

to employ digital technologies beyond content moderation.[107] These dissemination techniques aim to match users with content and with other users, in such a way that maximizes the user engagement with the platform.[108] Examples of dissemination techniques include content ranking, by deciding which content users should see first on their feed, as well as content recommendation, by suggesting groups, other users, and specific content. Hence, this work employs the term "content management" to refer to content hosting, content moderation, content ranking, and content recommendation.

The digital technologies utilized by online platforms are referred to as algorithms i.e., automated programs following a set of instructions designed to produce a certain outcome.[109] Thus, when referring to content management algorithms, the conceptualization employed in this thesis encompasses three types of algorithms *i.e.,* content moderation, ranking, and recommendation algorithms. This distinction is primarily relevant in the context of Chapter 5.

1.5 Outline of the study

This thesis is composed of six Chapters, including this introduction (Chapter 1) and a conclusion (Chapter 6). Chapters 2 to 5 result from four independent articles, each corresponding to one substantive Chapter and addressing one Research Question, and were published in peer-reviewed academic journals as sole or first author.[110] Chapter 5 is submitted for review in a peer-reviewed academic journal. As these Chapters were originally published as independent articles all contributing to the same problem statement and all following a

107 *E.g.* Rachel Griffin (2023). The Law and Political Economy of Online Visibility: Market Justice in the Digital Services Act. *Technology & Regulation,* 2023, 69-79. available at <https://doi.org/10.26116/techreg.2023.007> accessed 28 August 2024; Paddy Leerssen (2020) "The soap box as a black box: Regulating transparency in social media recommender systems." *European Journal of Law and Technology* 11.2.

108 Paul M Di Gangi and Molly M. Wasko (2016) "Social media engagement theory: Exploring the influence of user engagement on social media usage." *Journal of Organizational and End User Computing* (JOEUC) 28.2: 53-73

109 Kate Crawford (2021) *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

110 Eva Nave (2022) "Hate Speech, Historical Oppressions, and European Human Rights." *Buff. Hum. Rts. L. Rev.* 29: 83; Eva Nave and Lottie Lane (2023) "Countering online hate speech: How does human rights due diligence impact terms of service?." *Computer Law & Security Review* 51: 105884; and Eva Nave, Stephan Raaijmakers, and Thijs Veugen (2024) "Disrupting violence while preserving encryption: A human rights approach." *Technology and Regulation*: 115-131. These articles correspond, respectively, to the first, second, and third articles presented in this thesis in the substantive Chapters II, III, and IV.

logical sequence and, there is some overlap in sections providing context.[111] This thesis follows the following outline:

Chapter 2: Legal conceptualization of hate speech – Hate speech, historical or systematic oppressions, and European human rights

Chapter 2 answers Research Question 1 by providing a working definition of hate speech based on a critical exploration of the interplay between historical or systematic oppressions and the applicable European human rights standards. The European human rights standards are derived from conceptualizations of hate speech at the Council of Europe and at the European Union levels. This Chapter identifies the need to clarify legal standards applicable in online platforms to counter criminal hate speech originating from the conceptualization of criminal hate speech in CM/Rec(2022)16.

From Chapters 3 to 5, this thesis aims to clarify and advance new regulatory approaches to strengthen the human rights responsibilities of online platforms in the European context to counter criminal hate speech. The corporate human rights responsibilities framework states that businesses should adopt and follow policies and processes to respect human rights, including: "(i) a policy commitment to respect human rights; (ii) a human rights due diligence (HRDD) process to identify, prevent, mitigate and account for adverse impacts on human rights; (iii) processes to enable the remediation of any human rights abuses."[112] Following this structure,[113] the second, third and fourth articles seek to clarify and advance new regulatory approaches to strengthen the human rights responsibilities of online platforms to, respectively, prevent (Chapter 3), mitigate (Chapter 4), and remediate (Chapter 6) criminal hate speech.

111 With regards to text repetition, with the exception of some text presenting verbatim excerpts of relevant frameworks which was repeated in Chapters III-VI (e.g. Paragraph 11 of the CM/Rec(2022)16), I have chosen to include in this thesis the articles as originally published.

112 UNGPs (n 82), Principle 15.

113 The structure adopted in this thesis regarding the responsibilities of online platforms draws a parallel and builds on the work by Tarlach McGonagle concerning the States' human rights obligations. McGonagle subdivides positive state obligations into three categories: preventive, promotional, and remedial obligations. The thesis applies a similar subdivision to the human rights responsibilities of online platforms; Tarlach McGonagle (2019) "The Council of Europe and Internet Intermediaries" Human Rights in the Age of Platforms: The MIT Press, 241. With this approach, this thesis also builds upon the corporate human rights due diligence cycle stemming from the UNGPs and the CSDDD.

Chapter 3: Human rights responsibilities of online platforms to prevent criminal hate speech – How do European corporate preventive human rights responsibilities impact terms of service?

Chapter 3 answers Research Question 2 by investigating how the preventive human rights responsibilities of online platforms impact the design of terms of service and, more specifically, how corporate preventative human rights responsibilities of online platforms impacts the conceptualization of criminal hate speech on terms of service. This Chapter advocates that European legislators at the Council of Europe and at the European Union could and should require online platforms in the European context to align the conceptualization of hate speech in their terms of service with the European human rights conceptualization of criminal hate speech.

Chapter 4: Human rights responsibilities of online platforms to mitigate criminal hate speech – Disrupting incitement to violence in large groups on end-to-end encrypted services in Europe

Chapter 4 answers Research Question 3 by exploring the human rights responsibility of online platforms providing end-to-end encryption (E2EE) services to mitigate criminal hate speech in the form of incitement to violence. This Chapter provides a human rights review of technological developments and approaches available for content moderation on E2EE services, including metadata, hashing, and homomorphic encryption. The suggested mitigation strategy for content moderation in this context is disruption of incitement to violence towards historically or systemically oppressed communities.

Chapter 5: Human rights responsibilities of online platforms to remediate criminal hate speech – A call for a thorough corporate remedial responsibilities framework in Europe for criminal hate speech attributable to online platforms

Chapter 5 answers Research Question 4 by reviewing the European legal framework on the right to remedy applicable to victims[114] of online criminal hate speech and examining how the current regulatory framework applicable to online platforms conceptualizes the remedial responsibilities for platforms which amplified criminal hate speech. This Chapter evaluates the challenges with the current framework and proposes standards for a more comprehensive approach to corporate remedial responsibilities of online platforms in the

114 This research recognizes the civil society arguments against legal expressions patronizing the agency of marginalized people and thus avoids the use of "victims" and "protected characteristics", and uses instead people targeted by hate speech. *E.g.* Jennifer L Dunn, "Victims" and "survivors": Emerging vocabularies of motive for "battered women who stay." Sociological inquiry 75, no. 1 (2005): 1-30.

European context aiming for a better reconciliation with the right to remedy of people targeted by criminal hate speech disseminated by platforms.

Chapter 4: Conclusion

Chapter 6 addresses the Problem Statement by distilling the key findings of Chapters 2-5, by answering the Research Questions, and by highlighting areas for further research. This Chapter concludes with recommendations for European legislators, both at the Council of Europe and at the European Union, for online platforms, and for law enforcement bodies.

2 Legal conceptualization of hate speech

Hate speech, historical or systematic oppressions, and European human rights[12]

ABSTRACT

Today, around 5 billion people communicate through the Internet. While the benefits of online communication are undeniable, we also witness the proliferation of online hate speech, often associated with an increase in offline violence. Internet intermediaries and public bodies have developed frameworks to counter online hate speech. However, current frameworks lack a standardized approach to the conceptualization of hate speech. Some conceptualizations are overbroad, and others are underinclusive; overbroad because they lead to the removal of legal content (e.g. removal tools deleting legal content posted by marginalized communities), and underinclusive as the context of posts by linguistic minorities is often disregarded. This Chapter proposes a new legal conceptualization of hate speech in the European context. It does so by analysing the European regulatory framework through the lens of the first legal conceptualizations of hate speech deriving from critical (race) theory and (black) feminist intersectionality theory. The European focus is justified by the need to standardize at the regional level the legal requirements in current and future policies to counter online hate speech. The methodology is doctrinal, normative, and interdisciplinary legal research. There are two main findings. First, this Chapter suggests that the European regulatory framework needs to explicitly acknowledge the conceptualization of hate speech by critical legal scholars as expressions intended to perpetuate historical or systematic oppression. Second, this Chapter advocates that the conceptualization of hate

1 This Chapter was originally published in the Buffalo Human Rights Law Review 29(2022/2023): 83-145.

2 This Chapter was updated after publication and hence some content deviates from what was previously published. More specifically, references to legal and policy frameworks were updated to reflect the latest available information. Examples include 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 (Digital Services Act, DSA), the Regulation of the European Parliament and of the Council Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act, AI Act), and the Directive of the European Parliament and of the Council on combating violence against women and domestic violence. Cross-references should be read as referring to other references within the present Chapter. The citation style in this Chapter deviates from the standard European citation because it was originally published in an American Law Review.

speech in the European context can only achieve legal cohesion when all European regulatory instruments expressly account for the intersectionality of systems of oppression.

2.1 INTRODUCTION

The Internet enables borderless communications for more than half of the world's population. It connects people who are physically apart and it facilitates the spread of ideas and information. While the benefits of the Internet are undeniable, it also presents a dark side: hateful speech, for instance, tends to spread much faster and farther online,[3] often systematically targeting marginalized groups.[4] As noted by a former Secretary-General of the United Nations, the use of the Internet to promote hateful expressions is one of the most significant human rights challenges arising from technological developments.[5]

Online platforms have been linked to the rise of hate speech and violent conduct. For example, Facebook was accused of contributing to anti-Muslim riots in Sri Lanka and of playing a crucial role by hosting commentary inciting to violence against the Rohingya minority in Myanmar.[6] Other platforms have been associated with mass shootings, e.g. in the cases of Gab in relation to the Pittsburgh synagogue mass shooter and of 8kun with the El Paso killing.[7]

---

3 Binny Mathew et al., *Spread of Hate Speech in Online Social Media*, Proceedings of the 10th ACM Conference on Web Science 173 (2019), https://doi.org/10.1145/3292522.3326034 (last visited Oct 1, 2022).

4 This research acknowledges the ongoing debate about the use of the word "victim" as it may be interpreted to mean that those targeted are in a passive state of subjugation. This research acknowledges also the growing advocacy, especially by members of the civil society, for the use of the word "survivor" as it contains an emphasis on the strength of the people targeted by hate speech. However, it is also debated how this emphasis on strength may burden the targeted community with the obligation to overcome such traumatic experiences. With this background discussion in mind, this research will opt for using the expression "people targeted by hate speech" as much as possible. Nevertheless, the word "victim" may sometimes be used simply for legal coherence as this is the word used in the European Union Victims' Rights Directive 2012/29/EU.

5 U.N. Secretary-General, *Globalization and Its Impact on the Full Enjoyment of All Human Rights"* ¶¶ 26–28, U.N. Doc. A/55/342 (Aug. 31, 2000).

6 Michael Safi, *Sri Lanka Accuses Facebook Over Hate Speech After Deadly Riots*, Guardian (Mar. 14, 2018), https://www.theguardian.com/ world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech; Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. Times (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html.

7 *See* Lizzie Dearden, *Gab: Inside the Social Network Where Alleged Pitts burgh Synagogue Shooter Posted Final Message*, The Independent: Tech (Oct. 28, 2018, 8:10 PM), https://www.independent.co.uk/tech/pittsburgh-synagogue-shooter-gab-robert-bowers-final-posts-online-comments-a8605721.html; Tim Arango, Nicholas Bogel-Burroughs & Katie Benner, *Minutes*

In reaction to these events and due to international pressure, online platforms have started to self-regulate hate speech. However, such self-regulatory efforts often lack a standardized approach to the conceptualization of hate speech that is aligned with human rights. Though some online platforms expressly prohibit hate speech (e.g. Facebook, Twitter, YouTube, LinkedIn, TikTok, Tumblr, Microsoft), they then differ in their definitions. While Facebook defines hate speech 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,"[8] others do not refer directly to hate speech and focus instead on the prohibition of expressions based on their harmful impact (e.g. Reddit, WhatsApp, LinkedIn).[9]

A specific example of how the platforms' definitions of hate speech may not be aligned with human rights standards is Facebook's definition of "*protected categories*." In 2017, controversies arose when Facebook employed a definition of the categories protected from hate speech, disregarding the protections assigned under international and European human rights law to marginalized groups. In this case, that definition led to the removal of a post suggesting that "*all* white people were racist" but authorized a post incentivizing the "killing of *radicalized* Muslims." This decision was based on the justification that "radicalized Muslims" was a subgroup of a protected marker (i.e. religion), while "all whites" was more generic, thus supposedly more impactful, and therefore deemed more important to protect.[10]

Automated content moderation tools have been said to often be either overbroad or underinclusive. Overbroad because they take down content with no legal basis for removal (online hate speech detection tools have been under scrutiny for racial and queer[11] biases), and underinclusive as they often disregard context or content shared by linguistically marginalized groups.

Before El Paso Killing, Hate Filled Manifesto Appears Online, N.Y. Times (Aug. 3, 2019), https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html.

8 *Facebook Community Standards: Hate Speech*, META TRANSPARENCY CENTER, https://transparency.fb.com/en-gb/policies/community-standar ds/hate-speech/ (Jan. 13, 2023).

9 *See Content Policy*, REDDIT, https://www.redditinc.com/policies/con tent-policy#:~:text=Abide%20by%20community%20rules.,with%20or%20 disrupt%20Reddit%20communities.&text=Respect%20the%20privacy%20of%20others; *WhatsApp Terms of Service*, WHATS APP, https://www.whats app.com/legal/terms-of-service/?lang=en; *Professional Community* Policies, LINKEDIN, https://www.linkedin.com/legal/professional-community-policies.

10 Julia Angwin, ProPublica & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017, 5:00 AM), https:// www.propublica.org/article/facebook-hate-speech-censorship-internal-do cuments-algorithms.

11 This research uses the term "queer" as an umbrella term to refer to all LGBTIQ+ people, though it acknowledges the ongoing discussion that using the full acronym can be beneficial as a statement to expressly recognize the historically most marginalized groups within the queer community.

To date, there is no legally binding definition of *hate speech* in European or international human rights law. States and public bodies have passed legislation regulating online hate speech but controversies arise on how to conceptualize hate speech and on how to design effective legislation compliant with human rights standards.

The main questions that this Chapter seeks to answer are: (1) what are the main elements of hate speech under European human rights law?, (2) do they align with the original conceptualization of hate speech by critical legal theory?, and (3) to what extent do they require further clarification? By addressing these questions, this Chapter aims to clarify the main aspects of a legal conceptualization of hate speech, grounded in critical legal theory, laying the foundation for an analysis of advances and shortcomings in the European regulatory framework. The focus is on the European context as there is a need to systematize at the regional level the legal requirements for current and future hate speech policies.

The methodology is composed of doctrinal, normative, and interdisciplinary legal research. Doctrinal research focusing on applicable legal frameworks to online hate speech in Europe will contribute to clarifying the existing legal standards. Normative research will identify and address legal loopholes. Interdisciplinary legal research will investigate the interplay between European human rights law and critical legal (race) theory and (black) feminist intersectionality theory. These last two theoretical frameworks were selected as the term (racist) "hate speech" was coined and conceptualized within these fields.

Section 2.2 explores the legal foundations of what hate speech is, what its consequences are, and how it should be regulated from a critical legal perspective. The original legal conceptualization of racist hate speech by critical race theory is key to understanding that hate speech is used against historically and systematically oppressed groups. The insights by critical legal theory also help to understand the impact and harm of hate speech by highlighting the cumulative effects of continued exposure to hate speech and the intersectionality of systems of oppression (race,[12] gender, sexual orientation, etc.). This Section explores the legal foundations of the regulation of hate speech in three different periods: from the Enlightenment, passing by the 1980s and the insights from critical race theory, and to present times. This Section highlights how freedom of expression was never understood as an absolute right and how, since the start of the debate about systematic marginalization, exceptions to free speech have always been accounted for. It concludes with an analysis of the current legal challenges related to hate speech. For instance, the need

12 This research rejects theories of different human "races" as all humans belonging to the same species. However, this research refers to "race" or "racialized" groups as a means to expose a colonial and imperial process whereby a dominant group ascribes to another group a racial identity for the purpose of continued exclusion and domination.

to grant protection to people increasingly targeted by misogynistic and queerphobic hate speech, as well as hate speech targeting people with disabilities. Another current challenge relates to the digitalization of hate speech and how the legal system now needs to account for a faster and further dissemination of hate speech through the internet.

Section 2.3 investigates the theoretical underpinnings of hate speech at the Council of Europe.[13] This Section focuses both on treaty and non-treaty initiatives. The primary treaty is the European Convention of Human Rights (ECtHR), analysed together with relevant case law by the European Court of Human Rights. Other treaties are the Additional Protocol to the Convention on Cybercrime, concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems, the Convention on preventing and combating violence against women and domestic violence, and the Framework Convention for the protection of national minorities. Non-treaty initiatives selected for this analysis include: recommendations and guidelines by the Committee of Ministers; general policy recommendations by the European Commission against Racism and Intolerance (ECRI); outcomes of the European Ministerial Conferences on Mass Media Policy; outcomes of the Council of Europe Conferences of Ministers responsible for media and new communication services; and the Venice Commission Report on the relationship between freedom of expression and freedom of religion.[14] Section 2.3.4.1. analyses the main non-treaty framework which is the Recommendation CM/Rec(2022)16 of the Committee of Ministers to Member States on combating hate speech. This Recommendation draws on the main jurisprudence of the ECtHR on hate speech and is a cornerstone in the clarification of the main elements of hate speech in this Chapter.

Section 2.4 explores the main elements of hate speech in the substantive regulation at the European Union (EU) level. This Section starts by examining the EU's general principles and primary sources such as the Treaty of the EU and the Charter of Fundamental Rights of the EU. It then explains the main advances in the regulation of hate speech in secondary sources of the EU law

13 This Part's structure builds on the work of Tarlach McGonagle, *The Council of Europe Against Online Hate Speech: Conundrums and Challenges*, 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).

14 The Parliamentary Assembly of the Council of Europe (PACE), one of the two statutory organs of the Council of Europe, has also been an important actor working to counter hate speech, *e.g.* Resolution 1510 (2006), Recommendation 1805 (2007), Resolution 1743 (2010) and also through the No Hate Parliamentary Alliance. United Nations, Office of the High Commissioner for Human Rights (2011) Relevant Council of Europe Standards and Policies on the Prohibition and Prevention of "Hate Speech", Prepared by Directorate General of Human Rights and Legal Affairs (DGHL), available at <https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/ICCPR/Others2011/CouncilofEurope.pdf> accessed 21 November 2024. Nevertheless, this work prioritizes the analysis of Recommendations by the Committee of Ministers as the decision-making body of the Council of Europe.

such as: the Council Framework Decision on combating certain forms and expression of racism and xenophobia by means of criminal law; the Audiovisual Media Services Directive; resolutions adopted by the European Parliament (EP); the Regulation of the EP and of the Council on a Single Market for Digital Services (Digital Services Act, DSA);[15] the Regulation of the EP and of the Council Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act, AI Act); and Directive (EU) 2024/1385 of the EP and of the Council on combating violence against women and domestic violence.[16] Finally, this Section will explore the EC communication from December 2021 on its intention to extend the list of EU crimes to include hate speech and hate crime. In doing so, it does not focus on the procedural regulation of online hate speech as that pertains to the corporate human rights due diligence responsibilities of internet intermediaries moderating illegal content. Rather, the scope of this Chapter focuses on the substantive conceptualization of hate speech.

Section 2.5 concludes with a summary of the main elements in the legal conceptualization of hate speech rooted in European human rights law and supported by notions of critical theory and intersectionality advanced by the (black) feminism scholarship. Though the main elements in the conceptualization of hate speech were clarified in CM/Rec(2022)16, this Chapter presents two main findings. First, it is critical that the European regulatory framework explicitly acknowledges the scholarship of critical legal scholars in that they conceptualized hate speech as expressions intended to perpetuate historical or systematic oppressions. Second, this Chapter advocates that the conceptualization of hate speech in the European context can only achieve legal cohesion when all European regulatory instruments expressly account for the intersectionality of systems of oppression.

2.2 Legal Theoretical Foundations of Hate Speech

To date, there is no legally binding definition of *hate speech* in European or international human rights law. As noted by McGonagle, hate speech has been a "term of convenience" as it is often used to refer to a wide range of extremely negative content which would otherwise be very difficult to refer to.[17] For

15 The DSA also amends Directive 2000/31/EC, also referred to as the e-Commerce Directive.

16 European Commission, Directive (EU) 2024/1385 of the European Parliament and of the Council of 14 May 2024 on combating violence against women and domestic violence, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AL_202401385> accessed 21 November 2024.

17 Tarlach McGonagle, Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas 317 (2011). *See also* Podcast series by Katie Pentney "Decoding Hate," Episode 2 "The Hate You Tweet" with Tarlach McGonagle, 10 February 2021, funded by OSCE Representative on Freedom of the Media, #SAIFE project, available at <https://www.

example, the term "hate speech" has been informally employed to cover a wide range of situations from incitement to violence, hatred, bias, prejudice, insults, or defamation.[18]

This Section explores the legal foundations of hate speech by focusing on the framework developed by critical legal theory, and more particularly, critical race theory (CRT). The framework provided by CRT is a key starting point in this Chapter for the conceptualization of hate speech, as this was the scholarship in the 1990s that coined the term referring to "racist hate speech." CRT is a derivative body of critical legal theory. Critical legal theory, as per Young's definition, is a field of research that analyses society through its historical and sociological contexts.[19] CRT builds on these premises to challenge ahistoricism and to underline how current social and institutional inequalities derive from periods where racist intentions and practices were clearly outspoken.[20] This Section applies critical legal thinking to explain the first legal conceptualization of hate speech, the impact and harm it causes, the foundational legal debates on how to protect fundamental rights, and, current challenges to regulation.

2.2.1 First Legal Conceptualization

The concept of hate speech was coined by the legal scholarship *critical race theory* (CRT) in the early nineties in the United States of America in reference to "racist hate speech."[21] Following a 1990 report by the National Institute Against Prejudice and Violence that highlighted high levels of ethno-violence toward minority students on campuses, many universities and public bodies adopted regulations prohibiting speech stigmatizing racial minorities and other historically subordinated groups. These regulatory initiatives limiting racist expression sparked significant debate in American society and they were met with a strong sentiment of appreciation by members of the communities

decodinghatepod.com/episodes/episode-05-the-anywhere-.w orkout-lhgdz-D0JBu> accessed 4 October 2021; Tarlach McGonagle, *The Council of Europe Against Online Hate Speech: Conundrums and Challenges*, *in* Freedom of Expression and Democracy in the Digital Age: Opportunities, Rights, Responsibilities 40, 44 (2013).

18 James B. Jacobs & Kimberly Potter, Hate Crimes: Criminal Law and Identity Politics 11 (2001).

19 Iris Marion Young, Justice and the Politics of Difference 5 (2011).

20 Mari J. Matsuda et al., Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment 6 (2018); Derrick Bell, And We Are Not Saved: The Elusive Quest for Racial Justice (2008).

21 Matsuda et al. (n 20), 1.

targeted by such hateful speech as well as strong resentment by free speech First Amendment absolutists.[22]

The critical race legal scholars, many being racialized and themselves targets of hate speech, underlined the urgency to halt "racially abusive hate speech" and advocated for the restriction of freedom of expression in cases of racist speech.[23] CRT challenges ahistoricism and underlines how current social and institutional inequalities derive from periods where racist intentions and practices were clearly outspoken.[24] CRT aims to inspire legal and political systems that are informed about the victims' lived experiences.[25] More specifically, this scholarship seeks to provide evidence for how hate speech negatively affects the victim's rights, including dignity, non-discrimination, equality, participation in public life, expression, association, and religion, as well as how hate speech has the potential to inhibit self-fulfilment, self-esteem, and inflict physical harm.[26]

Matsuda, a critical race scholar who helped first conceptualize racist hate speech, contends that racist speech should not be treated as protected discourse under the First Amendment because it is the continuation of subjugation of groups historically oppressed.[27] Matsuda advocates for the prosecution of the worst forms of hate speech to provide public redress for the most serious harm. She proposes three elements to support the identification of the worst forms of 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.[28] This Chapter explores how Matsuda's conceptualization of hate speech applies to times when digital technology is pervasive and where people are not only being targeted with hate speech on the basis of their race, but also gender, sex, sexual orientation, disabilities, and age.[29]

22 U.S. Const. Amend. I ("Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances").
23 Matsuda et al. (n 20), 2.
24 Matsuda et al. (n 20), 6.
25 Matsuda et al. (n 20), 6; Jean Stefancic & Richard Delgado, Critical race theory: The Cutting Edge (2000); Richard Delgado & Jean Stefancic, Understanding Words That Wound (2019).
26 Matsuda et al. (n 20), 25.
27 *Id.* at 35.
28 *Id.* at 36
29 *Id*. at 16.

2.2.2 Impact and Harm

Along with the progress achieved by CRT on the conceptualization of racist hate speech, CRT was also the first line of scholarship to formally identify the harm caused to people when targeted by hate speech. Delgado and Stefancic highlight physical, psychological, and economic harm caused to people targeted by hate speech.[30] The short-term physical harms vary from rapid breathing, headaches, raised blood pressure, dizziness, and rapid pulse rate. In fact, scientists suspect that the high blood pressure and higher deaths from hypertension, hypertensive disease, and stroke from which many African Americans suffer could be associated with greater racial discrimination.[31] Potential long-term physical harm may in the worst cases lead to hate crime.[32]

With regard to psychological harms, victims of hate speech may experience fear, nightmares, low self-esteem, withdrawal from society (they forego their own right to freedom of expression), post-traumatic stress disorder, psychosis, anger, depression, and rejection of identification with their own race.[33] These effects have a different impact depending on the age of the victim. For example, children are believed to be among the most easily damaged by racial name-calling.[34] This negative impact on youngsters can be heightened when parents experience discriminatory practices themselves and have to put energy into overcoming their own trauma while educating their children.[35] Finally, psychological harms can also lead to self-harmful behaviours.[36]

People targeted by hate speech also experience economic harms. For example, research has shown that racialized students at white-dominated universities may earn lower grades as a result of stress caused by the continuous exposure to racist behaviour.[37] Similarly, racialized people who manage to succeed academically and professionally are often in white-dominated environments and, thus, are more likely to encounter racism. As explained by Matsuda, such experiences may lead hate speech victims to change jobs, forgo education, avoid public spaces and restrict their own freedom of expression to avoid receiving hate messages.[38]

In assessing the harm caused to the targets of hate speech, another key element from CRT that this Chapter explores is the notion of *intersectionality*.

---

30 Delgado & Stefancic (n 25), 12–19.
31 Delgado & Stefancic (n 25), 13.
32 Delgado & Stefancic (n 25), 5.
33 Delgado & Stefancic (n 25), 14.
34 Joe R. Feagin & Debra Van Ausdale, The First R: How Children Learn Race and Racism (2001).
35 Joe R. Feagin & Debra Van Ausdale (n 34).
36 A. Sumner et al., *Association of Online Risk Factors with Subsequent Youth Suicide-Related Behaviors in the US*, JAMA Network Open (Sept. 20, 2021), https://jamanetwork.com/journals/jamanetworkop en/fullarticle/2784337Steven'.
37 Stefancic & Delgado (n 25), 111–121.
38 Matsuda et al. (n 20), 24.

Intersectionality is an analytical framework that explains how elements of someone's social and political identities merge to create different forms of discrimination or privilege. Although the call for attention to intersectionality dates back to Black feminists from the 19th century, when Cooper questioned the overlap of women and race questions,[39] Crenshaw introduced the concept in legal scholarship in 1989.[40] She proposed intersectionality as both a metaphor for crossing categories of discrimination and as a means to show the shortcoming of approaches that seek to isolate systems of oppression.[41] She warned about how the politics and hateful discourse of race and gender have worked to exclude and marginalize especially racialized women.

Crenshaw further suggested the legal concepts of discrimination need to be revised if they are intended to serve as remedies to historical or systematic oppression.[42] Notably, even though the primary intersections explored by the intersectionality theory were race and gender, Crenshaw clarified that the concept could and should be expanded by considering characteristics such as class, sexual orientation, and age.[43] To help clarify the contextual systems of oppression, it is valuable to highlight how Matsuda insists that legal scholars should listen to the victims of hate speech and recognize their entitlement to directly express their concerns and fears; Matsuda claimed that victims would only find redress through such a representation process.[44] This expansive consideration of elements contained in the research of intersectional systems of oppression is essential in the analysis developed further in this Chapter.

Hate speech also harms the perpetrator and society as a whole. For instance, as the perpetrator of hate speech deepens their hateful beliefs, they can develop a paranoid mentality with respect to the community that they routinely denigrate, leading to the spread of hateful beliefs within the community of the perpetrator.[45] Finally, hate speech impacts society altogether as it challenges the fundamental value of equality, and equal respect and dignity, security and the rule of law.[46] In fact, social scientists have shown how people

39 Anna Julia Cooper, A Voice from the South 21 (Zenia, The Aldine Printing House 1892); Anna Carastathis, Intersectionality: Origins, Contestations, Horizons 15 (University of Nebraska Press 2016).

40 Kimberlé Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, 1989 U. CHI. LEGAL F. 139, 141 (1989).

41 Kimberle Crenshaw, *Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Colour* (1991), Stanford Law Review, Vol. 43, No. 6 (Jul., 1991), pp. 1241-1299.

42 Crenshaw (n 40), 140.

43 Crenshaw (n 41), 1244-45.

44 Matsuda et al. (n 20), 114.

45 Matsuda et al. (n 20), 24.

46 Jeremy Waldron, The Harm in Hate Speech 92 (2012).

targeted by hate crimes take much longer to recover compared to people targeted by crimes without racist motivation.[47]

Finally, this Chapter agrees with critical race theorists in that it is important to consider the cumulative effect of continued exposure to hate speech. More specifically, groups of people who have been historically targeted by hate speech may develop emotions of self-hatred, especially when the victim internalizes the negative perception of who they are.[48] The various *types* of harm caused by hate speech are real and require legal redress to avoid perpetuating historical oppressions. As advocated by Parekh, hate speech should be restricted both "for what it is and for what it does."[49] In establishing a legal framework for countering hate speech, it is important to understand how the regulation of hate speech interplays with other rights and to question if there are different *degrees* of hate speech requiring different legal courses of action.

2.2.3 Regulation and Balancing Conflicting Rights

In the previous subsections 2.2.1. and 2.2.2, this Chapter explained the original conceptualization of (racist) hate speech and its impacts as presented by critical race theory. The following paragraphs explore the foundational debates on the regulation of hate speech; more specifically, the debates on the balance of competing rights when regulating hateful expression. As any form of regulation of hate speech inevitably affects the exercise of the right to freedom of expression, this subsection explores how legal scholars have balanced competing rights – freedom of speech versus dignity[50] and prohibition of discrimination.[51] This analysis will focus on two different periods: first, the Age of Enlightenment, and second, the period from the 1990s marking the first conceptualization of racist hate speech by critical race theory.

Analysing the origins of the right to free speech and its original conceptualization during the Age of Enlightenment, it is possible to conclude that, at its conception, free speech was not considered an absolute right in situations where it inflicted *serious harm* to others. The right to freedom of expression

47 Frederick M. Lawrence, *The Punishment of Hate: Toward a Normative Theory of Bias-Motivated Crimes*, 93 MICH. L. REV. 320, 342–343 (1994).

48 Richard Delgado, *Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling*, 17 HARV. C. R.-C. L. L. REV. 133, 137 (1982); Charles R. Lawrence, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 Stan. L. Rev. 317, 201 (1987).

49 Bhikhu Parekh, *Is There a Case for Banning Hate Speech?* in The Content and Context of Hate Speech: Rethinking Regulation and Responses 37 (Michael Herz & Peter Molnar, eds., 2012).

50 The right to human dignity is contained in Art. 1 of the EU Charter of Fundamental Rights.

51 The right to freedom from harm is herewith used to refer broadly to the right to dignity and the right to non-discrimination.

was a prominent debate in the Liberalist ideals from the 17th and 18th centuries.[52] Liberalism prioritizes individual freedoms such as freedom of expression, thought, and association. This period was marked by an emphasis on the importance of preserving the marketplace of ideas, in which the truth would only emerge from a free, transparent and public competition of ideas.[53] However, even the most renowned liberal thinkers did not advocate for an absolutist conception of freedom of expression. In fact, Locke proposed a non-absolutist notion of tolerance when he defended opinions contrary to human society, or the moral rules necessary to the preservation of society, should not be legally protected.[54] Mill went further to propose legal restrictions on freedom of expression particularly when it caused "harm to others."[55]

Despite the existence of some considerations of the non-absolutist notion of freedom of speech at the beginning of Liberalism, it was in the 1990s that the legal debate about historical oppressions and the restriction of racist hate speech ignited in the United States. There are two key elements in the debate on the regulation of hate speech from the beginning of critical race theory to the present times. First, there were new stronger arguments for the protection of marginalized communities which were challenged by traditional constitutionalist US scholars. Second, there was more formal discussion about the harm in hate speech as the continuation of historical oppression.

First, critical legal scholars challenged the Liberalist conception of equal opportunities to exercise rights and underlined the need to restrict hateful speech, as it perpetuated harmful practices toward the already marginalized community.[56] As noted by Waldron, perpetrators of hate and degradation intend to undermine dignity and destabilize the identity attributes shared by members of targeted oppressed communities.[57] Also, as proposed by MacKinnon, "a well-ordered society" must assure that members of marginalized groups live in a dignified manner.[58]

However, they were met with strong opposition by other scholars from the United States Constitutional tradition. For example, Dworkin stated that no individual should be shielded from content that may negatively impact

52 Liberalism was a societal system that placed the emphasis on individual liberty, equality, and consent to be governed, and supported access to common resources in proportion to the individual's contribution. Liberal ideals replaced feudalism, which established a static set of social classes with different fixed tasks and different fixed possibilities to access resources, indifferent to an individual's contribution to the common good.

53 Caitlin Ring Carlson, Hate Speech 10 (2021).

54 John Locke, A Letter Concerning Toleration (Merchant Books 2011) (1765).

55 John Stuart Mill, On Liberty and Other Essays (John Gray ed., Oxford Univ. Press 1998) (1859).

56 *See Section* 2.1 and 2.2.

57 Waldron (n 46), 92.

58 Catharine A. MacKinnon, Only words (1993).

their self-esteem.[59] Scanlon emphasized stronger protection for the speaker, pluralism and non-interference, often at the expense of the rights of the group targeted by that hate speech, despite admitting that such speech can cause harm.[60] Traditional American scholars' resistance to the regulation of hate speech is flawed because there were already many situations limiting freedom of expression,[61] and such an absolutist interpretation of the First Amendment's response to hate speech simply perpetuates racism.[62] Additionally, these traditional constitutionalist American scholars fail to recognize the limits in their positionality,[63] as they fail to recognize their position of power and privilege and their likely (un)conscious bias. As advocated by Cohen-Almagor, the American viewpoint of neutrality defended by Scanlon is inherently false for it fails to recognize the inequalities and discriminatory practices in access and exercise of basic freedoms.[64]

Second, hate speech is a term that has been used informally across many fields such as law, sociology, criminology, and psychology, to refer to a vast number of harmful expressions including incitement to hatred, insults, defamation, bias, and prejudice. All these expressions could happen simultaneously or separately and could respectively cause different degrees of harm. This Chapter aligns with Post's claim that for an expression to be hate speech it needs to be "extreme" in nature because no legal order can aim to abolish emotions of intolerance and dislike.[65] However, it is herewith proposed a broader interpretation of extreme because hate speech is in fact, in and of itself, harmful expression. There is nevertheless a distinction to be made between hate speech and the most serious forms of hate speech which should have

59 Ronald Myles Dworkin, Freedom's Law: The Moral Reading of the American Constitution 260 (1996).

60 See Thomas Scanlon, *Freedom of Expression and Categories of Expression Principles of Expression and Restriction: A First Amendment Symposium*, 40 U. PITT. L. REV. 519, 527 (1979). Nevertheless, it should be noted that, in a more recent work published in 2018, Scanlon acknowledged that insult and harassment causing psychological harm might be grounds for exclusion of expression, *e.g.* Thomas Scanlon (2018) A Framework for Thinking about Freedom of Speech and some of its Implications, available at <https://www.law.berkeley.edu/wp-content/uploads/2018/10/Freedom-of-Speech-Berkeley.pdf> accessed 21 November 2024. See *also* Raphael Cohen-Almagor, *Racism and Hate Speech – A Critique of Scanlon's Contractual Theory*, 53 FIRST AMENDMENT STUDIES 1, 2 (2019).

61 Examples include privacy, individual reputation, protection of intellectual property, regulation of economic markets, speech infringing public order. *See* Delgado & Stefancic (n 25), 11, 34; Frederick Schauer, *Categories and the First Amendment: A Play in Three Acts*, 34 Vand. L. Rev. 265, 270 (1981).

62 Schauer (n 61), 270; Parekh (n 49), 222.

63 Positionality is the concept that someone's personal experiences of race, gender, class, etc., and location in time and space affect one's view of the world. Issues of positionality refute ideas of neutrality, and objective research shows that any researcher will inevitably design and conduct research in a subjective way influenced by their positionality in the world.

64 Cohen-Almagor (n 60), 22.

65 McGonagle (n 17), 14 (citing Ivan Hare & James Weinstein, Extreme Speech and Democracy 123 (2009)).

implication in the actionable legal area i.e. civil and administrative law for hate speech and criminal law for the most serious forms of hate speech.

In assessing whether an expression is extreme and causes serious harm, States should investigate the context in which the expression was manifested and should specifically investigate if the targeted community has been historically oppressed. In examining if the targets of hate speech have been historically oppressed, as pointed out by Boyle and Baldaccini, it is important to investigate the "core mischiefs" of hate speech i.e. the impact of hate speech on the exercise of other rights.[66] Examples of this exercise would be to look for how the targets of hate speech have been accessing among others their rights to dignity, to have a house, to have a job, and to education by investigating the activities and claims by civil rights movements.

2.2.4 Current Legal Challenges

The critical legal theory, foundational in the conceptualization of hate speech explained in Section2.1, has been particularly challenged by various recent developments in terms of impact as well as mediums, and reach. First, the impact of hate speech is now broader. Recent data shows that, aside from the racialized community and from women, hate speech now also affects more seriously LGBTIQ+ people, people with disabilities, and members of religious groups.[67] The expanded recognition of the groups targeted by hate speech calls for a more formal legal acknowledgment of the intersectionality aspects of the victims and the present regulation does not currently reflect such development, often still being solely applicable for cases of racist hate speech.

Second, there are also new regulatory challenges related to the mediums and reach of hate speech since the beginning of the digitalization era. With the advent of the Internet, hate speech is spreading faster and further online,[68] potentially also leading to an increase in offline violence.[69] Additionally, the online spaces where hate speech has been spreading are mostly run by U.S.-based companies operating on the basis of American traditional constitutionalism, i.e., the First Amendment which predicates on an almost absolutist interpretation of freedom of expression. The current dominance of U.S.-based companies running online platforms operating worldwide has created difficulties

66 *Id.* at 319 (citing Kevin Boyle & Anneliese Baldaccini, A Critical Evaluation of International Human Rights Approaches to Racism 152 (2017)).

67 *The EU Code of Conduct on Countering Illegal Hate Speech Online*, European Commission, https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en (last visited March 12, 2023).

68 Mathew et al. (n 3).

69 Karsten Müller & Carlo Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crime*, 19 J. EUR. ECON. ASS'N 2131 (2021).

in a decentralized public and for contextual oversight of the hate speech rules guiding our freedom of expression and right to non-discrimination online.

Third, the fast spread of harmful content online has led to the use of the term "hate speech" by various scholarships in a much broader way than in legal scholarship. At present times, hate speech continues to be a term *informally* used across many research fields to cover a spectrum of expressions including: (a) criminal offenses; (b) problematic expressions, which although not criminal offenses, could be prohibited under civil or administrative law; and, (c) content which bears no legal implications but still raises issues of respect and tolerance.[70]

The following Sections 2.3 and 2.4 will first analyse respectively how hate speech is conceptualized at the Council of Europe and at the European Union and seek to clarify how these regulatory systems align with the critical legal foundation discussed in Section 2.2. They will also address the current regulatory challenges with respect to online hate speech as explained in this subsection.

2.3 Approaches to Hate Speech at the Council of Europe

2.3.1 General Objectives

There is no explicit mention of hate speech in the European Convention for the Protection of Human Rights and Fundamental Freedom (ECHR) or in any other treaties of the Council of Europe (the CoE or The Council). Still, the CoE developed various instruments which are useful in the regulation of hate speech. All these instruments follow the main objectives of the Council which are to uphold human rights, democracy, tolerance, non-discrimination and the rule of law.[71]

Nevertheless, despite this alignment in terms of general legal principles, the various regulatory initiatives by the Council to address hate speech pursue different specific goals and contain different procedural and substantive thresholds. As a result, there is a significant variety of strategies ranging from legal action countering hate speech, to education, and promotion of increased representation of minorities in the media.

---

70 U.N. High Commissioner for Human Rights, Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred, ¶12, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013).

71 Statute of the Council of Europe preamble, May 5, 1949, 87 U.N.T.S. 103.

2.3.2 ECHR and ECtHR Jurisprudence

This Section focuses on European human rights law governing hate speech by addressing, firstly, the most relevant provisions in the European Convention of Human Rights (ECHR) and subsequently, investigating the jurisprudence of the European Court of Human Rights (ECtHR) in the application of such provisions.

The ECHR does not contain a direct reference to hate speech. Instead, it regulates other rights impacting the regulation of hate speech. These rights are often in opposition and their application requires legal reasoning to understand how to draw the balance between competing human rights.

From the perspective of the speaker of potentially hateful content, the most relevant article is the right to freedom of expression and to what extent there can be restrictions on its exercise (Art. 10(2)). From the perspective of the people targeted by potentially hateful speech, the most relevant provision is the prohibition of abuse of rights (Art. 17). Additionally, the ECHR also prescribes the prohibition of discrimination (Art. 14), later expanded in Protocol 12,[72] and grants the right to respect for private life (Art. 8). To fully grasp the legal reasoning to strike the balance between competing rights, it is important to look into the jurisprudence of the ECtHR.

The first reference of the ECtHR to hate speech dates from 1999,[73] though the term was not developed at the time. Still, and despite the fact that the Court never provided a concrete and fixed definition of hate speech, it developed the meaning of hate speech in various instances thereafter. The ECtHR has generally declared that statements attacking or casting in a negative way an entire group on the basis of for example ethnicity, race, and religion, are in contradiction with the underlying values of tolerance, social peace and non-discrimination prescribed by the ECHR.[74] Yet, as the conceptualization of the harm caused by hate speech continues to be very context-dependent, it is crucial to clarify the legal strategies deployed by the Court in cases potentially amounting to hate speech to understand the applicable European human rights standards.

Before explaining the legal strategies in more detail, it is important to point out some of the key interpretative principles guiding the Court in the effort

72 Protocol 12 to the European Convention on Human Rights art. 1, Nov. 4, 2000, E.T.S. 177.

73 Drawing on the work of McGonagle, the term hate speech was first used in four judgments of the ECtHR, all of July, 8, 1999: Sürek v. Turkey (No. 1), App. No. 26682/95, ¶ 62 (July 8, 1999), https://hudoc.echr.coe.int/eng?i=001-58279; Sürek & Özdemir v. Turkey, App. Nos. 23927/94 & 24277/94, ¶ 63 (July 8, 1999), https://hudoc.echr.coe.int/eng?i=001-58278; Sürek v. Turkey (No. 4), App. No. 24762/94, ¶ 60 (July 8, 1999), https://hudoc.echr.coe.int/eng?i=001-58298; Erdogdu & Ince v. Turkey, App Nos. 25067/94 & 25068/94, ¶ 54 (July 8, 1999), https://hudoc.echr.coe.int/eng?i=001-58275; McGonagle (n 17), 11.

74 Perinçek v. Switzerland, App. No. 27510/08, ¶ 206 (Oct. 15, 2015), https://hudoc.echr.coe.int/eng?i=001-158235.

to balance competing rights in the ECHR, especially when balancing the right to freedom of expression (Art. 10) and the right to non-discrimination (Art. 14) or right to private life (Art. 8).[75] First, as freedom of expression represents such a cornerstone of the protection of democracy and the rule of law, the Court decides on a case-by-case basis regarding cases on restrictions of freedom of expression.

Second, the Court follows the margin of appreciation doctrine[76] whereby it assigns a certain discretion to national courts in the domestic interpretation and application of ECHR provisions, subject to the Court's supervisory jurisdiction role. The Court has considered that States are better placed to appreciate the meaning of public morals, decency and religion as these may vary considerably from country to country. Thus, instead of having a cross-cutting understanding of the meaning of such concepts, in reviewing domestic cases impacted by such considerations, the Court has sought to assess whether the justifications by national authorities are relevant and sufficient.

Third, the Court defends that the rights prescribed in the ECHR must be "practical and effective."[77] This means that rights in the ECHR must not be interpreted in an elusive and hypothetical way and, as a consequence, the violation of rights in the ECHR should lead to effective remedial procedures (Art. 13).

Fourth, the Court considers the ECHR a "living instrument" which indicates that the rights contained in the ECHR "must be interpreted in the light of present-day conditions."[78] As the living conditions change, rights can too be challenged by unforeseen circumstances and, as such, the ECHR must continue to be interpreted in such a way that continues to grant protection of rights even if applied in different future contexts unanticipated at the time of the drafting.

Finally, the Court developed a positive obligations' doctrine whereby it requires that States ensure every person can exercise the rights in the ECHR

75 E.g. in Kaboğlu v. Turkey, the Court found that there had been a violation of Article 8 emphasizing that the negative impact of the verbal attacks and threats of physical harm against the applicants. Kaboğlu v. Turkey, App. Nos. 1759/08, 50766/10 and 50782/10 (October 30, 2018). In Beizaras and Levickas v. Lithuania, the Court found that there had been a violation of Article 14 in conjunction with Article 8, concluding that the applicants had been discriminated on the grounds of their sexual orientation. Beizaras and Levickas v. Lithuania, Case Number 41288/15 (January 14, 2020). In a similar case, the Court found that the Romanian authorities had failed to discharge their positive obligation to investigate if a verbal homophobic abuse had amounted to a criminal offence. In this case, the Court found a violation of both Articles 14 and 8; see Association ACCEPT and Others v. Romania, Application no. 19237/16 (June 1, 2021).

76 The margin of appreciation doctrine was included in the Preamble to the ECHR with the adoption of the Convention Amending Protocol No. 15 in August 2021.

77 Airey v. Ireland, App. No. 6289/73, ¶ 24 (Oct. 9, 1979).

78 Tyrer v. the United Kingdom, App. No. 5856/72, ¶ 31 (April 25, 1978); Matthews v. the United Kingdom, App. No. 24833/94, ¶ 39 (Feb. 18, 1999).

not only through a non-interference principle but also when necessary through interfering and measures protecting the exercise of rights.[79]

Turning to the legal strategies applied by the Court when ruling on cases concerning hate speech, as explained by former Judge Tulkens, these can be summarized in two possible approaches.[80] In the first approach (explained in Section 2.3.2.1), the ECtHR can apply Art. 17 (prohibition of abuse of rights) and exclude protection of said hateful expression from the ECHR because such conduct violates or limits (to an extent greater than the one provided in the ECHR) any right in the ECHR, and it is therefore considered an abuse of rights. In the second approach (explained in Section 2.3.2.2), the ECtHR can apply Art. 10(2) when the hateful expression is not considered to violate or limit fundamental rights in the ECHR, but it could still amount to a hateful expression that should be restricted if it meets the conditions in Art. 10(2) (explained in more detail in Section 2.3.2.2).

2.3.2.1 Hate Speech as a Clear Abuse of Rights

As per the jurisprudence of the ECtHR in its first approach to hate speech, hate speech can trigger the prohibition of abuse of rights provision (Art. 17 ECHR). Before elaborating on the substance of the case law related to the application of Art. 17 to hate speech cases, a note is necessary on the procedural consequences of the application of the prohibition of abuse of rights.

The ECtHR uses Art. 17 mainly when it considers whether there is a clear abuse of rights because the act *in casu* either violates or limits (further than what is allowed in the ECHR) rights in the ECHR. In such cases, the Court deems the application of a case inadmissible on its merits. Such declaration of inadmissibility means that, as the case represents such a serious abuse of rights, the Court refuses to proceed with the process of balancing of rights and will not proceed to the judgment on the substance. In essence, when Art. 17 is invoked for hate speech cases, it means that these are the most blatant cases of hate speech. Therefore, the Court dismisses such cases as manifestly ill-founded[81] and does not even proceed to the assessment of Art. 10(2) on whether the restriction to the right to freedom of expression of the speaker of hate speech was rightfully applied.[82]

Elaborating on the case law related to the substantive application of the prohibition of abuse of rights (Art. 17), i.e., to the conceptualization of the most

79 Özgür Gündem v. Turkey, App. No. 23144/93, ¶ 43 (March 16, 2000), https://hudoc.echr.coe.int/eng?i=001-58508.

80 Françoise Tulkens, The Hate Factor in Political Speech: Where Do Responsibilities Lie?, Report of the Council of Europe Conference, Warsaw, Sept. 18-19, 2013.

81 ECHR Art. 35(3).

82 Seurot v. France, App. No. 57383/00 (May 18, 2004), https://hudoc.echr.coe.int/fre?i=002-4404. *See also* Factsheet – Hate Speech, September 2020, Press Unit, European Court of Human Rights, 1-5.

serious cases of hate speech, the Court has ruled that these include: negationist and revisionism; incitement to discrimination, hatred or violence on the grounds of race, ethnicity, religion or sexual orientation likely to give rise to feelings of rejection and hostility; threats to democracy by expressions inspired by totalitarian views; and support to terrorism. Each of these forms of the most serious cases of hate speech will now be explained.

First, negationist and revisionist practices negating the Holocaust or other genocides[83] or other crimes against humanities should be considered hate speech and a clear abuse of rights under Article 17. The Court has supported that such expressions amounted to some of the most serious examples of hate speech because they denied clearly established historical facts that cannot be considered historical research and cannot be considered an endeavour pursuing the truth; had the purpose of rehabilitating the National Socialist regime; double victimized the victims accusing them of falsifying history.[84]

Second, acts that incite discrimination, hatred or violence and are likely to give rise to feelings of rejection and hostility have also been considered by the ECtHR as amounting to the most serious cases of hate speech and deemed a clear abuse of rights under Article 17. For example, in Le Pen v. France,[85] the Court agreed with the national court's conviction of incitement to discrimination, hatred and violence toward a group of people because of their origin or membership of non-membership of a specific ethnic group, nation, race or religion. The ECtHR declared the comments made by the president of the French National Front in an interview saying that "the day that there are no longer 5 million but 25 million Muslims in France, they will be in charge" indeed represented Muslim community as a whole in a disturbing light, likely to give rise to feeling of rejection and hostility, and therefore, interference had been necessary to protect the democratic values.

In assessing such cases of hate speech inciting discrimination, hatred or violence, this Chapter suggests the interpretation of the protected categories should be open-ended. Even though the ECtHR has formally recognized the

83 It is noteworthy that the First Additional Protocol to the Cybercrime Convention is the first international treaty that applies to genocides other than the Holocaust (Art 6).

84 Garaudy v. France, App. No. 65831/01, ¶ 23 (March 25, 2003), https://hudoc.echr.coe.int/fre?i=002-4830. Other examples include M'bala M'bala v. France, App. No. 25239/13 (Oct. 20, 2015), https://hudoc.echr.c oe.int/eng?i=001-160358; Honsik v. Austria, App. No. 25062/94 (Oct. 18, 1995), https://hudoc.echr.coe.int/eng?i=001-2362; Marais v. France, App. No. 31159/96 (June 24, 1996), https://hudoc.echr.coe.int/eng?i=001-88275; Williamson v. Germany, App. No. 64496/17 (Jan. 8, 2019), https://hudoc.echr.coe.int/eng?i=001-189777; Pastors v. Germany, App. No. 55225/14 (Jan. 3, 2020), https://hudoc.echr.coe.int/eng?i=001-196148; Peta Deutschland v. Germany, App. No. 43481/09 (Mar. 18, 2013), https://hudoc.echr.coe.int/eng?i=001-114273; Perinçek v. Switzerland, App. No. 27510/08 (Oct. 15, 2015), https://hudoc.echr.coe.int/eng?i=001-139724.

85 Le Pen v. France, App. No. 18788/09, (Apr 20, 2010).

protection against hate speech discriminating only on the basis of the race,[86] religion,[87] and ethnicity,[88] other grounds should also be regarded as impermissible grounds. In fact, the Court recalled in Lilliendahl v. Iceland that discrimination based on sexual orientation is as serious as discrimination based on "race, origin or colour" and underlined the importance to grant protection from hateful and discriminatory speech to groups on the basis of gender and sexual minorities and stressed the historic marginalization that these groups have endured.[89]

An important element in the application of this category of hate speech (incitement to discrimination, hatred or violence on the grounds of race, ethnicity, religion or sexual orientation likely to give rise to feelings of rejection and hostility) is that it requires a negative generalization of the targeted group. For example, in Norwood v. the UK, the ECtHR specified that linking a (religious) group as a whole with a grave act of terrorism was incompatible with the democratic values of tolerance, social peace and non-discrimination.[90]

---

86 *See, e.g.*, Glimmerveen & Hagenbeek v. Netherlands, App. No. 8348/78 (Oct. 11, 1979); Medya FM Reha Radvo ve Iletisim Hizmetleri A. S. v. Turkey, App. No. 32842/02 (Nov. 14, 2006); Simunic v. Croatia, App. No. 20373/17, (Jan 22, 2019).

87 *See, e.g.*, Belkacem v. Belgium, App. No. 343667/14, ¶ 21 (Jun. 27, 20 17), https://hudoc.echr.coe.int/fre?i=001-175941; Gunduz v. Turkey, Ap p. No. 35071/91, ¶18 (Nov. 13, 2020), https://hudoc.echr.coe.int/fre?i=001-6 1522.

88 *See e.g.*, Ivanov v. Russia, App. No. 35222/04, ¶ 10 (Feb. 20, 2007); W.P. & Others v. Poland, App. No. 42264/98, ¶ 47 (Sept. 2, 2004), https://hudoc.echr.coe.int/fre?i=001-66711.

89 Lilliendahl v. Iceland, App. No. 29297/18, ¶ 45 (May 12, 2020) (Nevertheless, this is a controversial case, and some remarks are necessary. On the one hand, this is a remarkable case as the Court confirmed the protection of the queer community from hate speech and, more specifically, the Court sided with the domestic court's view that the homophobic online comment amounted to hate speech. On the other hand, though the Court declared the case inadmissible and manifestly ill-founded, it did not consider the comment as one of the most serious forms of hate speech. The Court stopped short of granting more solid protection to the queer community when it declared that the applicant's comment did not amount to the most serious forms of hate speech claiming that it was not clear the aim was to incite violence and hatred or to destroy rights in the ECHR. This position fails to align with the criteria used to investigate the most serious cases of hate speech. Establishing a parallel with the Le Pen v. France case, some key elements in the comment should have been sufficient for the EctHR to recognize incitement to discrimination and hatred in a disturbing way likely to give rise to a feeling of rejection and hostility, and therefore constitute an example of the most serious cases of hate speech. These include the fact that the comment clearly stated that gay people had a sexual deviation, such said sexual deviation was characterized by derogatory words mainly used to describe animals, and finally the fact that the applicant considered homosexual people disgusting and repulsed their representation in the media. These statements seriously undermine the broader reach of the impact of hate speech online and compromise the right to respect for private life, non-discrimination, and to freedom of information, as they specially legitimize oppression of a historically repressed group. The Lilliendahl v. Iceland case is essential to clarify that homophobic speech is hate speech but cannot be considered as a deviation from the jurisprudence regarding the criteria used to investigate the most serious cases of hate speech.)

90 Norwood v. United Kingdom, App. No. 23131/03, ¶ 13 (Nov 16, 2004).

Similarly, in Pavel Ivanov v. Russia, the ECtHR also declared that "accusing an entire ethnic group of plotting a conspiracy" was a general, vehement attack on one ethnic group and is directed against the Convention's underlying values of tolerance, social peace and non-discrimination.[91]

In assessing the prohibition of hate speech in the form of incitement to discrimination, hatred or violence on the grounds of religion, the Court specified that only cases of criminal offenses should be included under Article 17.[92] More specifically, in Mariya Alekhina and Others v. Russia, the ECtHR ruled that insults to religious beliefs and blasphemy should not be subject to criminal sanctions.[93] This represents a narrower affordance of domestic autonomy within the margin of appreciation doctrine.

Third, another category of hate speech considered serious enough to amount to a prohibition of abuse of rights is speech that threatens the democratic order because it is inspired by totalitarian views. As a rule, the ECtHR will declare inadmissible applications inspired by totalitarian doctrines or which express ideas that represent a threat to the democratic order and are liable to lead to restoration of a totalitarian regime.[94]

Fourth, the Court has also interpreted expressions supporting terrorist activities as serious cases of hate speech under Article 17. In Roj TV A/S v. Denmark, the media service had applied against its conviction for terrorism offenses by Danish courts for promoting the Kurdistan Workers' Party (PKK) through television programs. The ECtHR found that the PKK could be considered a terrorist organization within the meaning of the Danish Penal Code and supported the domestic conviction in light of the margin of appreciation doctrine. However, this case could be read as to dismiss an important debate on the lack of due process in the European Union for the classification of the PKK as a terrorist organization.[95] Therefore, questions arise in this case as to whether the Court conducted an effective investigation to protect media independence and access to information. Moreover, given the ECtHR's interference with the margin of appreciation doctrine in the case Mariya Alekhina

91 *See e.g.*, Ivanov v Russia, App. No. 35222/04, ¶ 20.

92 *Id.* This position confers a better alignment of the ECtHR with international human rights standards on the balance exercise between the restrictions to freedom of expression and freedom of religion.

93 *See, e.g.*, Alekhina v. Russia, App. No. 22519/02, ¶ 224 (July 13, 2006), https://hudoc.echr.coe.int/fre?i=001-73321.

94 *See, e.g.*, Communist Party of Germany v. Federal Republic of Germany, App. No. 250/57 (July 20, 1957), https://hudoc.echr.coe.int/eng?i=001-110191; B.H., M.W., H.P. & G.K. v. Austria, App. No. 12774/87 (Oct. 12, 1989), https://hudoc.echr.coe.int/eng?i=001-1039; Nachtmann v. Austria, App. No. 36773/97 (Sept. 9, 1998), https://hudoc.echr.coe.int/eng?i=001-4399; Schimanek v. Austria, App. No. 32307 (Feb. 1, 2000), https://hudoc.echr.coe.int/eng?i=001-24075

95 *European Court: Decisions Placing the PKK on the List of Terrorist Organizations Annulled*, PRAKKEN D'OLIVEIRA (Nov. 15, 2018), https://www.prakkendoliveira.nl/en/news/2018/european-court-decisions-placing-the-pkk-on-the-list-of-terrorist-organizations-annulled.

and Others v. Russia in the same year as Roj TV A/S v. Denmark, it could be argued that also in the latter case the Court could have at least acknowledged the controversies on the classification of PKK as a terrorist group.

Recapitulating, the application of Article 17 leads the Court to dismiss the case as manifestly ill-founded, thus in principle not even leading to an assessment of the balance of rights. Nevertheless, there were three cases in which the Court, despite considering the application inadmissible, still provided further details as to why specific cases of hate speech were amongst the most serious hateful expressions. First, in Šimunić v. Croatia,[96] a case of offline hate speech, the ECtHR sided with the domestic courts stressing the need to tackle racism and totalitarian ideas shared by prominent sports figures.

Second, in Smajić v. Bosnia and Herzegovina,[97] a case regarding online hate speech, the Court found the domestic courts had rightfully convicted the applicant for national, racial, and religious hatred, discord or intolerance following posts on an internet forum describing military action that could be undertaken against Serb villages. The ECtHR ruled that the penalties of a suspended sentence and a seized laptop had not been excessive and dismissed the claim as manifestly ill-founded.

Third, in Nix v. Germany,[98] also a case regarding online hate speech, the Court sided with the domestic court's decision to consider the applicant criminally liable for an online post of a picture of a Nazi leader and a swastika because it had not been clear the applicant's intent to reject the Nazi ideology, finding the application manifestly ill-founded. In this case, the Court reiterated that Article 10 of ECHR applied to the Internet, as did the conditions for the restriction of freedom of expression in Article 10(2). The ECtHR added that States which have experienced the Nazi horrors may be regarded as having a special moral responsibility to distance themselves from the Nazi ideology in light of their historical role and experience from the mass atrocities perpetrated by the Nazis.[99] This case could be read as to assign special moral responsibilities of States with history of totalitarian regimes to counter online hate speech.

In summary, the Court applies Article 17 in cases of criminal hate speech, which in its view should only be applied to the most serious cases of hate speech. When analysing the jurisprudence of the ECtHR on Article 17 in the light of the original conceptualization by Matsuda[100] of hate speech as ex-

96 Šimunić v. Croatia, App. No. 20373/17 (Jan. 22, 2019), https://hudoc.echr.coe.int/eng?i=001-189769.

97 Smajić v. Bosnia & Herzegovina, App. No. 48657/16 (Jan. 18, 2018), https://hudoc.echr.coe.int/eng?i=001-180956.

98 Nix v. Germany, App. No. 38285/16 (Mar. 13, 2018), https://hudoc.echr.coe.int/eng?i=001-182241.

99 *Id.* ¶ 47.

100 The conceptualization of (racist) hate speech introduced by Matsuda is explained on Section 2.2.1. of this Chapter.

pressions perpetuating historical or systematic oppressions, it is possible to conclude the Court fails to consistently mention historical or systematic oppressions as a key element of hate speech and fails equally to recognize the intersectionality of systems of oppression perpetuated by hate speech.

2.3.2.2 No Clear Abuse But Hate Speech is Prohibited

Under the second approach, i.e., in cases where there is no clear abuse of rights as per Article 17 ECHR and where the application is not considered inadmissible on its merits, the ECtHR applies the right to freedom of expression (Art. 10).[101] An essential point of departure is that the Court posited that freedom of expression applies "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."[102]

Any restriction on the right to freedom of expression speech needs to follow the criteria contained in Article 10(2). That is, the Court considers a restriction of freedom of expression to be legal if it is: (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.

When assessing the necessity criterion, the ECtHR evaluates whether there is a pressing social need, whether the restriction is proportional and the relevance and sufficiency of States' justifications. It is relevant to highlight that for restrictions to be deemed necessary, they need to aim at protecting the interests of national security, public safety, the prevention of disorder or crime, the protection of health or morals and the protection of the rights and freedoms of others.[103] In this evaluation, the ECtHR leaves to the States a margin of appreciation, recognizing the diversity in social and legal traditions across all Member States.[104]

---

101 It is noteworthy that although Article 10 (right to freedom of expression) has been the main point of departure in the ECtHR jurisprudence, hate speech can impact a variety of other rights such as non-discrimination, life, association, etc.

102 Handyside v. UK, App. No. 5493/72, ¶ 49 (Dec. 7, 1976), https://hudoc.echr.coe.int/eng?i=001-57499.

103 For a detail study, see *e.g.* Janneke Gerads (2013) "How to improve the necessity test of the European Court of Human Rights." *International journal of constitutional law* 11.2: 466-490.

104 The margin of appreciation doctrine was used first in Lawless v. Ireland, E.Ct.H.R. Ser.B, 1960-61, para.90, p.82. Nevertheless, this doctrine's prominence was more definitively affirmed in Handyside v. U.K. in 1976. For a detailed analysis see Hutchinson, M. R. (1999) The Margin of Appreciation Doctrine in the European Court of Human Rights. *The International and comparative law quarterly*. [Online] 48 (3), 638–650.

When assessing the severity of the hateful expression, the ECtHR has developed a set of standards which Rosenfeld describes as the "contextual variables"[105] approach. Generally, this approach has considered the victims' perspectives;[106] political and social background;[107] the intent of the speaker;[108] the speaker's status or role in society;[109] the content of the expression;[110] the reach of the expression;[111] and, the nature of the audience.[112] With regard to the element of intentionality, this Chapter aligns with the dissenting opinion by Judges Ryssdal, Bernhardt, Spielman and Loizou in Jersild v. Denmark in that the presumably good intentions of the person disseminating hate speech are not enough to dismiss hate speech when such hateful expressions do provoke racist statements (in this case, the journalist's intentions). In fact, critical race scholars stress the importance of emphasizing the impact and harm over the speaker's potentially non-existent intention to discriminate.[113] Regarding the element of the speaker's status, the Court has deemed it particularly relevant as it has ruled that politicians,[114] teachers,[115]

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

106 *See, e.g.*, Leroy v. France, App. No. 36109/03, ¶ 27, 31, 43 (Oct. 2, 2008), https://hudoc.echr.coe.int/eng-press?i=003-2501837-2699727.

107 *See, e.g.*, *id.*; Ceylan v. Turkey [GC], App. No. 23556/94 (July 8, 19-99), https://hudoc.echr.coe.int/fre?i=002-6560; Beizaras & Levickas v Lit huania, App. No. 41288/15 (Jan. 14, 2020), http://hudoc.echr.coe.int/eng? i=001-200344.

108 *See, e.g.*, Jersild v. Denmark, App. No. 15890/89 (July 8, 1993).

109 *See, e.g.*, Incal v. Turkey, App. No. 22678/93 (June 9, 1998), https://hudoc.echr.coe.int/fre?i=001-58197 (noting politicians enjoy a protected status, but concomitantly have heightened responsibilities in that they should avoid disseminating comments in their public speeches which are likely to foster intolerance); Féret v. Belgium, App. No. 15615/07 (July 16, 2009), https://hudoc.echr.coe.int/eng-press?i=003-2800730-3069797 (noting that politicians have the duty to refrain from using or advocating for racial discrimination).

110 *See, e.g.*, Goucha v. Portugal, App. No. 70434/12 (Mar. 22, 2016), https://hudoc.echr.coe.int/fre?i=001-161527; Feldek v. Slovakia, App. No. 29032/95 (October 12, 2001), https://hudoc.echr.coe.int/fre?i=001-59588; Ottan v. France, App. No. 41841/12 (July 19, 2018), https://hudoc. echr.coe.int/fre?i=001-182627.

111 *See, e.g.*, Gündüz v. Turkey, App. No. 35071/97 (Dec. 4, 2003), https://hudoc.echr.coe.int/fre?i=001-61522 (stating that live TV as not easy to reformulate or retract).

112 *See, e.g.*, Vejdeland & Others v. Sweden, App. No. 1813/07 (May 9, 2012), https://hudoc.echr.coe.int/eng?i=001-109046; Vereinigung Bildender Künstler v. Austria, App. No. 68354/01 (April 25, 2007), https:// hudoc.echr.coe.int/fre?i=001-79213.

113 "Good intentions are not enough," as mentioned in Henrika McCoy, *Black Lives Matter, and Yes, You Are Racist: The Parallelism of the Twentieth and Twenty-First Centuries*, 37 CHILD ADOLESC. SOC. WORK J. 463, 464 (2020), https://www.degruyter.com/document/doi/10.7208/97802 26703725/html (last visited Oct 4, 2022) (citing ANNE WARFIELD RAWLS & WAVERLY DUCK, TACIT RACISM 90 (2020)).

114 Féret v. Belgium, App. No. 15615/07 (July 16, 2009). However, in Le Pen v. France the Court did not refer to the political status of the applicant.

115 *See, e.g.*, Lilliendahl v. Iceland, App. No. 29297/18 (June 12, 2018), available at <https://hudoc.echr.coe.int/fre?i=001-203199> accessed at 21 November 2024.

and famous sporters[116] have a higher responsibility not to engage in hate speech statements. As to the reach of hateful expressions, the ECtHR has not yet directly mentioned the increased reach of online hate speech as shown by recent studies.[117]

In general terms, the ECtHR's case law on expressions that should not be protected under Article 10 because it would constitute hate speech include:

1 content that encourages violence, armed resistance or insurrection if there is (i) an intentional and direct use of wording to incite to violence and (ii) where there is a real possibility that the violence occurs;[118]
2 "glorification of terrorism" when provoking public reaction (i.e., adherence from general public to the idea), which would be capable of stirring up violence and of having a demonstrable impact on public order;[119]
3 content that creates a "pressing social need"[120] to provide protection against hateful content and, to this, Judge Tulkens suggested that the ECtHR built its argument considering that the harmful effect depended on historic, demographic and cultural contexts;[121]
4 content of racist and xenophobic nature during electoral campaigns and pronounced by members of parliament;[122]
5 serious and prejudicial allegations, even if not a direct call to hateful acts.[123]

However, the interpretation of the pressing social need in conjunction with the margin of appreciation doctrine has been controversially applied in

116 *See, e.g.*, Šimunić v. Croatia, App. No. 20373/17 (Mar. 9, 2017), https://hudoc.echr.coe.int/fre?i=001-189769.
117 Mathew et al. (n 3).
118 *See, e.g.*, Erbakan v. Turkey, App. No. 59405/00 (July 6, 2006) (finding there had been a violation of Article 10 because there was no proof of actual risk or imminent danger of the speech fomenting intolerance); Vajnai v. Hungary, App. No. 6061/10 (Sept. 23, 2014) (finding the Government failed to adduce any evidence to suggest that there is real and present danger of any political movement or party restoring the communist dictatorship); Kiraly & Domotor v. Hungary, App. No. 10851/13 (April 17, 2017) (finding authorities failed to act against racial violence and they had breached the right to respect for private life under Article 8 ECHR); Beizaras & Levickas v. Lithuania, App. No. 41288/15, ¶ 128 (Jan. 14, 2020), http://hudoc.echr.coe.int/eng?i=001-200344 (finding authorities failed to prosecute as the comments on Facebook had constituted "undisguised calls on attacks on the applicants' physical and mental integrity, which required protection by criminal law"). *E.g.*, Dicle v. Turkey, App. No. 46733/99, ¶ 33 (July 11, 2006), https://hudoc.echr.coe.int/eng?i=001-223703.
119 *See, e.g.*, Leroy v. France (n 106).
120 *See, e.g.*, I.A. v. Turkey, App. No. 42571/98 (Sep. 13, 2005), https://hudoc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=001-70 113&filename=001-70113.pdf&TID=hcoelbxhnm.
121 *See, e.g.*, Soulas & Others v. France, App. No. 15948/03, ¶¶ 32-35 (Oct. 7, 2008), http://hudoc.echr.coe.int/eng?i=001-87370.
122 *See, e.g.*, Féret v. Belgium, App. No. 15615/07 (July 16, 2009), http://hudoc.echr.coe.int/eng?i=001-93626.
123 *See, e.g.*, Vejdeland & Others v. Sweden, App. No. 1813/07 (Feb. 9, 2012), https://hudoc.echr.coe.int/eng?i=001-109046.

Perinçek v. Switzerland,[124] when the Court declared a violation of Article 10 defending that in this particular context, hate speech had a diminished impact. The ECtHR looked into geographical, historical and temporal elements in the contextualization of hate speech to defend public statements in Switzerland calling the Armenian genocide a "lie" should not have been criminalized. In this judgment, as mentioned by Bayer and Bard, the Court departs from its case law in various aspects.[125]

First, the negation of genocide war crimes is to be covered as serious hate speech under Article 17 ECHR. The Court had consistently interpreted the cases of denial of the Holocaust genocide as inadmissible under Article 17 of the ECHR. Nevertheless, here the ECtHR found that the criminal conviction of the Armenian genocide denial statements was neither a "pressing social need" nor "necessary in a democratic society." This raises the question as to whether the Court would apply different thresholds for different genocides. Second, the Court fails to recognize the potential international reach of such revisionist statements. Third, this Chapter goes beyond the analysis by Bayer and Bard by also claiming this case deviates from the ECtHR jurisprudence that public figures have a special duty to not express hateful speech. *In casu*, the statements were made by a famous political figure which should have resulted in the application of the special duty to refrain from expressing hateful opinions. As a result, this case cannot but be analysed as a deviation from the European human rights standards set by the Court itself in its previous case law. Importantly, the politicians' duty to counter hate speech, including their duty to remove hateful contents posted by third-parties on the politician's social media accounts, was confirmed in a more recent ECtHR case.[126]

2.3.3 Other Treaties

Besides the ECHR, the CoE adopted other treaties that complement the regulation of hate speech. The following paragraphs present a selection of instruments impactful for the regulation of hate speech because they contain provisions on the right to non-discrimination and/or prohibitions of incitement to hatred. These treaties are introduced according to their descending order of relevance for the regulation of hate speech.

124 Perinçek v. Switzerland, App. No. 27510/08 (Oct. 15, 2015), https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22002-10930%22]}.

125 Judit Bayer & Petra Bard, Hate Speech And Hate Crime In The Eu And The Evaluation Of Online Content Regulation Approaches 38 (July 2020), Https://Www.Europarl.Europa.Eu/Regdata/Etudes/Stud/2020/ 655135/Ipolstu(2020)655135_EN.pdf.

126 Sanchez v. France, App. No. 45581/15 (ECHR, 15 May 2023), https://hudoc.echr.coe.int/eng#{%22itemid%22:[%22001-224928%22]}

The 2003 Additional Protocol to the Convention on Cybercrime (APCC)[127] was developed with the goal of harmonizing the criminalization of racist and xenophobic acts committed through computer systems. Such acts encompass the dissemination of racist or xenophobic material (Art. 3), racist and xenophobic motivated threats (Art. 4), insults (Art. 5), the denial, gross minimization, approval or justification of genocide or crimes against humanity (Art. 6). Aiding or abetting any of these conducts is also criminalized under this Protocol (Art 7). The APCC requires States to adopt and enforce legislation and/or other effective measures to make several types of racist conduct committed via computer systems *criminal* offences under domestic law "when committed intentionally and without right."

While representing a significant accomplishment in the protection against racism and xenophobia online, this Protocol raises further important legal debates. First, the inclusion of racist and xenophobic insults in Article 5 as a criminal act may be interpreted as to create a new standard when compared to the ECtHR jurisprudence on freedom of expression. This is because, although the ECtHR ruled expressions that "offend, shock or disturb" to be generally protected,[128] based on Article 5 of the Additional Protocol and in the event that "offensive expressions" and "insult" are interpreted as interchangeable terms, it can be argued that racist or xenophobic offenses spread through the internet can be criminally actionable. A rationale supporting this standard might be that expressions shared digitally typically reach a wider audience and this may have a far more impactful reach than when shared offline. This interpretation requires further discussion about the human rights safeguards protecting freedom of expression. Still, the acknowledgement of this heightened legal standard is a significant development in European human rights standards to counter racist and xenophobic motivated online hate speech.

Second, the APCC is not aligned with the European and international human rights standard that only the most serious cases of hate speech should be criminalized.[129] The focus on criminal law measures against online hate speech undermines the relevance of civil law or other non-legal responses as key strategies to counter and prevent further hate speech. Though the APCC provides for the possibility of not attaching criminal liability when other remedies are available and in case the conduct is not associated with hatred or violence (Art. 3(2)), this is only possible for acts of dissemination of online hate speech with racist or xenophobic intent. Furthermore, it seems unrealistic

127 Additional Protocol to the Convention on Cybercrime, Jan. 28, 2003, E.T.S. 189, https://rm.coe.int/168008160f.

128 Handyside v. UK (n 102) , ¶ 49.

129 *See, e.g.*, Eur. Ct. H.R. Press Release on Admissibility Decision of Le Pen v. France (App. No. 18788/09), and the recommendation of the Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression A/74/486 (May 7, 2010), https://hu doc.echr.coe.int/app/conversion/pdf/?library=ECHR&id=003-3117124-34 55760&filename=003-3117124-3455760.pdf.

to apply criminal law to most cases given the high prevalence of hate speech online and the well-known length of any legal action, even more so when it is criminal. Finally, this framework heavily impacted by the aforementioned challenges associated with the almost exclusive reliance on criminal law could be said to be in violation of Article 13 ECHR, for not contributing to an effective remedy for people targeted by hate speech. A revision of this standard seems to be necessary for full compliance with European human rights law.

Third, the restriction of this Protocol to hate speech with racist or xenophobic intent leaves out other hate speech such as misogynist and queerphobic speech. The restriction to racist or xenophobic content does not align with critical legal theory nor the theory of intersectionality as it dismisses many other systems of oppression. To conclude, despite representing a significant accomplishment in elevating the threshold of protection against racist and xenophobic insults online, the APCC contributes to an overuse of criminal law, to entropy in the remedial processes for victims, and leaves unprotected many groups targeted by online hate speech.

The 2011 Convention on preventing and combating violence against women and domestic violence (Istanbul Convention)[130] also contributes to the regulatory framework of (online) hate speech. The Istanbul Convention is key in guaranteeing protection from expressions manifested offline or online with the intent to threaten the target (Arts. 33 and 34) and from unwanted verbal or non-verbal expressions of a sexual nature manifested with the purpose or effect of violating the dignity of a person, in particular when creating an intimidating, hostile, degrading, humiliating or offensive environment (Art. 40). Aiding and abetting stalking is also considered as an offense (Art. 41).

Three remarks are relevant with respect to this instrument's contribution to the regulation of hate speech. First, even though Article 1 of the Convention initially refers to women and victims of domestic violence, the groups protected by the selected articles are to be broadly interpreted. Not only do these provisions refer to protecting people in general but the Explanatory report[131] clarifies the drafters' wish to suggest an open-ended list of grounds for non-discrimination.[132] Suggested grounds for the open-ended list of protected

130 Convention on Preventing and Combating Violence Against Women and Domestic Violence, May 11, 2011, E.T.S. 210, https://rm. coe.int/1680 08482e.

131 Though not binding, explanatory reports are important sources for the interpretation of international law instruments.

132 Explanatory Report, Convention on Preventing and Combating Violence Against Women and Domestic Violence, ¶ 53, May 11, 2011, E.T.S. 210, https://rm.coe.int/1680a48903. See also *id*. at ¶ 87 ("For the purpose of this Convention, persons made vulnerable by particular circumstances include: pregnant women and women with young children, persons with disabilities, including those with mental or cognitive impairments, persons living in rural or remote areas, substance abusers, prostitutes, persons of national or ethnic minority background, migrants – including undocumented migrants and refugees, gay men, lesbian women, bi-sexual and transgender persons as well as HIV-positive persons, homeless

categories include among others gender, sexual orientation, gender identity, age, state of health, disability, marital status, and migrant or refugee status. Gender is also broadly conceptualized. While the main text seems to have a narrow definition in Article 3 referring to male or female only, in the Explanatory report the drafters clarified they seek to protect people on the basis of gender in a more expansive meaning when they recognize that queer people may too be persecuted on the basis of their gender.[133] The conceptualization of victims of domestic violence is also broad and applies to *all* victims (Art. 2(2)).

Second, similarly to the APCC, the Istanbul Convention prescribes that expressions creating an offensive environment should not be protected (Art. 40). Again, this could arguably be read in contradiction to the ECtHR ruling that expressions that "offend, shock or disturb" are to be generally protected. The protected groups covered from hate speech under this provision should also be interpreted as an open-ended list. Nevertheless, in line with critical legal theorists, for the expression to amount to hate speech, a contextual analysis of potential historical oppressions would need to be conducted.

Third, also in the same way as the APCC, the Istanbul Convention seems to emphasize the preference for the use of criminal law for Articles 33, 34, and 40. It should be noted that other legal civil or administrative responses could be accepted as long as these provide other means of effective, proportionate, and dissuasive measures. Still, the emphasis on criminalization should always be expressly reserved to the most serious cases, and this caveat should have been better clarified in the Istanbul Convention.

Another treaty relevant to the regulation of hate speech is the 1994 Framework Convention for the Protection of National Minorities (FCNM).[134] The FCNM prohibits discrimination and calls for equal rights before the law,[135] and it encourages intercultural dialogue as well as measures to protect persons who are subject to threats or acts of discrimination, hostility, or violence due to their ethnic, cultural, linguistic, or religious identity.[136] This Framework Convention emphasises legal action to fight bias and the need to prevent hate speech, namely through empowering the role of the media.[137]

---

persons, children and the elderly"). This broad conceptualization is also supported in Art. 12 on the general preventive measures.

133 *Id.* ¶ 53.

134 Council of Europe, Framework Convention for the Protection of National Minorities and Explanatory Report, European Treaty Series – No. 157, Doc. H9510 (1995), available at <https://rm.coe.int/168007cdac> accessed 22 November 2024.

135 FCNM (n 134), Art. 4.

136 FCNM (n 134), Art. 6.

137 FCNM (n 134), Art. 9.

2.3.4 Non-Treaty Initiatives

There are also non-treaty instruments at the level of the Council of Europe, which contribute to defining the legal contours of hate speech in Europe. These instruments were selected on the basis of containing either direct references to hate speech or references to discrimination, tolerance, or to the protection of marginalized groups. Some instruments specifically regulate online harms, including online hate speech.

The study of the responsibilities of private actors when regulating online hate speech (procedural regulation) is not the main focus of this Chapter and will, therefore, not be developed in detail. Though some initiatives on business and human rights will be mentioned for ease of reference on instruments impacting the regulation of online hate speech, this Chapter focuses instead on clarifying the main elements in the conceptualization of hate speech (substantive regulation).

The non-treaty framework at the CoE is presented following the legal status of the sources and the subsequent order of influence of the instruments for the case-law of the ECtHR. Firstly, this Section will focus on the Recommendations and Guidelines of the Committee of Ministers (CM), and secondly, General Policy Recommendations of the European Commission against Racism and Intolerance (ECRI). The instruments produced by these bodies are the most important non-treaty initiatives because they are frequently mentioned by the ECtHR in its case law, and, as noted by McGonagle, the Court recognizes these bodies produce standard-setting work.[138] Subsequently, this Section will describe the most relevant communications at the CoE level countering hate speech.

The Committee of Ministers (CM) adopted three recommendations directly guiding the regulation of hate speech, two recommendations on measures to combat discrimination, and three recommendations with guidelines on how to regulate business and human rights, the latter particularly relevant for the regulation of online hate speech.

Starting with the recommendations directly guiding the regulation of hate speech, Recommendation No. R (97) 20 was the first direct and expansive communication from the CM on the topic of hate speech. This Recommendation defines hate speech as "all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, antisemitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin" and notes that governmental representatives must both refrain from hateful statements and establish a legal

138 As noted by McGonagle, the ECtHR recognized that both the CM and the ECRI manage to engage with specific topics in a more expansive manner while taking into consideration current practice and potential developments. *See* McGonagle (n 17), 27.

framework for civil, administrative and criminal responses.[139] The principles to combat hate speech stated in this recommendation place a significant emphasis on the need to have a legal framework where multiple stakeholders are responsible for contributing to the monitoring and countering of hate speech, highlighting the special responsibility of public officials and the media.[140]

Recommendation No. R (97) 21 was adopted immediately after Recommendation No. R (97) 20 and focuses on the need to capitalize on contributions by the media to prevent hate speech by promoting a culture of tolerance. This recommendation highlighted that, in order to counter hate speech, it was not enough to respond with legal measures and underscored the importance that the media adopted programs to promote access to media for minorities and codes of conduct promoting tolerance. An additional emphasis was placed on the importance to train media professionals on multiculturalism and to promote integration and airtime for all individuals, especially ensuring access and representation of marginalized communities.

On 20 May 2022, the CM adopted Recommendation CM/Rec(2022)16 in a wide-ranging strategy to combat hate speech in light of current challenges brought about by technological developments and the rise of hate speech prevalence, especially on social media. This recommendation was prepared by a Committee of Experts (ADI/MSI-DIS) that had been set up in 2020 by the CM with the goal of drafting an updated framework for a comprehensive human rights strategy to address hate speech, including in the online environment. The main findings and suggestions in CM/Rec(2022)16 build on existing CoE treaties and other standard-setting initiatives as well as on the ECtHR case law. This recommendation is a landmark instrument at the CoE level as it has the potential to provide a clear and updated roadmap for the regulation of hate speech in the broader European context. A dedicated reflection is included below in Section 2.3.4.1.

Two other CM recommendations address hate speech in the form of discrimination. Recommendation (2010)5 on measures to combat discrimination on grounds of sexual orientation or gender identity, expressly included the obligation to combat inciting hatred or other forms of discrimination against

139 Recommendation of the Committee of Ministers on Hate Speech, Doc. R 97 (1997).

140 Principle 1 specifies that public officials are under a special responsibility to refrain from stating or inciting hate, particularly in the media. Principles 2, 3 and 4 reinforce the idea that States must guarantee a legal framework composed of civil, administrative and criminal law to address hate speech, with the caveat that only the most serious hateful expressions should be criminalized. Additionally, in Principle 6, the CM differentiated between the responsibility of the author of hate speech and the responsibility of the media reporting such hate speech, underlining that the latter must be able to communicate on matters of public interest such as the abuse of freedom of expression through hate speech. Regarding Principle 6, this research argues though that this principle should not apply to online platforms with algorithm designed to promote hate speech given its virality. For further info, *see* McGonagle (n 17), 23.

the LGBTI+ community. In Recommendation (2011)7 on a new notion of media, the CM stated that digital platforms should monitor the use of biased expressions and defended that these actors be required by law to report, to the competent authorities, criminal threats of violence based on racial, ethnic, religious, gender or other grounds that come to their attention.

There are also three other CM recommendations on business and human rights with particular relevance for the regulation of online hate speech. Recommendation (2016)3 on human rights and business; Recommendation (2018)2 on the roles and responsibilities of internet intermediaries; and Recommendation (2020)1[141] on the impacts of algorithmic systems on human rights, together lay the groundwork for corporate responsibility to comply with human rights safeguards needed to prevent online hate speech through digital means.[142]

The CM has also in recent years produced two other instruments impacting the regulation of hate speech, which even if not in the form of recommendations, still reflect the views of the Committee and can thus influence the case-law of the ECtHR. In 2019, the CM adopted a Declaration on the manipulative capabilities of algorithmic processes[143] warning against the risk of using algorithmic processes to manipulate social and political behaviour. And in 2021, the CM adopted the Guidelines on upholding equality and protecting against discrimination and hate during times of crisis.[144] These guidelines alert the need to counter online hate speech during times of crises, namely by improving data collection on marginalized groups; by involving the affected community in this research; by providing continued access to information, legal, psychological and social support to victims; and by having public authorities as role models publicly rejecting hate speech. In a remarkable consideration, these guidelines also recognize the intersectional discrimination of hate speech.[145]

---

141 Committee of Ministers, Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems, Council of Europe, https://search.coe.int/cm/pages/result _details.aspx?objectid=09000016809e1154.

142 *Id.* Overall, the CM underscores the need to support national institutions in the oversight, risk assessment and enforcement of the United Nations Guiding Principles on Business and Human Rights by business running algorithmic systems.

143 Declaration by the Comm. of Ministers on the Manipulative Capabilities of Algorithmic Processes (Feb. 13, 2019), https://search. coe.int/cm/pages/result_details.aspx?ObjectId=090000168.

144 Steering Comm. on Anti-Discrimination, Diversity And Inclusion (CDADI), Guidelines of the Committee of Ministers of the Council of Europe on Upholding Equality and Protecting Against Discrimination and Hat e During the Covid-19 Pandemic and Similar Crises in the Future, Council of Europe (May 2020), https://policycommons.net/artifacts/1811064/guidelines-of-the-committee-of-ministers-of-the-council-of-europe-on-upholding-equality-and-protecting-against-discrimination-and-hate-during-the-covid-19-pandemic-and-similar-crises-in-the-future-2021/2547081/ on 05 Mar 2023. CID: 20.500.12592/0sjj6g.

145 CDADI (n 144).

The European Commission against Racism and Intolerance (ECRI) has also engaged specifically with the topic of hate speech. The most prominent General Policy Recommendation (GPR) is GPR No. 15 on combating hate speech, where ECRI defines hate speech as "advocacy, promotion or incitement, in any form, of the denigration, hatred or vilification of a person or group of persons, as well as any harassment, insult, negative stereotyping, stigmatization or threat in respect of such a person or group of persons and the justification of all the preceding types of expression, on the ground of 'race,' colour, descent, national or ethnic origin, age, disability, language, religion or belief, sex, gender, gender identity, sexual orientation and other personal characteristics or status." The use of the word "insult" for the typification of the hate speech acts, despite having to be firstly interpreted in light of national legislative provisions, could also raise questions of potential lack of coherence with the ECtHR Handyside judgment.[146] However, it is important to note that both the APCC and the Istanbul Convention already utilized respectively, racial and xenophobic insults and sexual harassment conducive to offensive environment as some of the most serious forms of hate speech, even suggesting their criminalization. Thus, one possible interpretation here is that insults addressed at historically oppressed groups on the basis of their race or gender (non-exclusive interpretation), would be considered a seriously grave form of hate speech. Notably, ECRI GPR 15 promotes self-regulation of media and states that criminalization is necessary following the Rabat Plan of Action,[147] i.e., in circumstances where hate speech is intended or can reasonably be expected to incite acts of violence, intimidation, hostility or discrimination against those targeted.

ECRI adopted three other GPRs relevant for this analysis. GPR 7 on national legislation to combat racism and racial discrimination contributed to remarkable developments for the regulation of hate speech when it declared that intent to incite the commission of acts of violence, intimidation, hostility or discrimination is not essential to criminalize. Instead, it indicated that criminal law can also be used when violence, intimidation, hostility or discrimination can reasonably be expected to be the effect of using the hate speech concerned and therefore the use of such speech would be considered reckless. GPR 6 on combating the dissemination of racist, xenophobic and antisemitic material via the Internet requested governments to take the necessary measures at national and international levels to act effectively against the use of the internet for racism, xenophobia, and antisemitism. GPR 11 on combating racism and racial discrimination in policing, ECRI asks States to ensure law enforcement investigates racist offenses in victim-friendly environments.

---

146 *See* Handyside v. UK (n 102). In the Handyside judgment the Court posited that freedom of expression applies also to ideas that offend, shock or disturb the State or any sector of the population.

147 Annual Report. of the U.N. High Commissioner for Human Rights: addendum, ¶ 34, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013).

Also relevant for the analysis of non-treaty initiatives on hate speech at the CoE is the 2008 Report by the European Commission for Democracy through Law (Venice Commission) on the relationship between freedom of expression and freedom of religion.[148] In this report, the Venice Commission defends incitement to hatred, including religious hatred, should have a specific requirement of the intention of recklessness to be criminalized and concludes the offense of blasphemy should be abolished.

Finally, the European Ministerial Conferences on Mass Media Policy and Council of Europe Conferences of Ministers responsible for Media and New Communication Services reiterate in various instances the need to tackle hate speech and promote tolerance. Of note, in 2013, the Ministers responsible for Media and Information Society decided "to protect people from the risks encountered on the Internet, in particular by fighting cybercrime, sexual abuse, exploitation of children, cyberbullying, gender-based discrimination, incitement to violence, hatred and any form of hate speech," and invited the CoE to "continue to combat hate speech and incitement to violence and terrorism, whether involving individuals, public or political persons or groups, including offering guidance on ways to mitigate its escalation, due to the speed and scope of its online dissemination."[149]

2.3.4.1 Observations Regarding the CM Recommendation on Combating Hate Speech

This Section provides a dedicated analysis of CM/Rec(2022)16[150] as this recommendation will hopefully pave the way for a comprehensive and renewed effort on the regulation of hate speech in the European context. For instance, this recommendation could influence the European Union's developing position on the regulation of hate speech – the European Commission communicated in December 2021 its intent to extend the list of the EU crimes to hate speech and hate crime.[151] This analysis will be divided into three main segments: (1) considerations related to the people targeted by hate speech, (2) considerations on the conceptualization of hate speech and (3) its regulation, and con-

148 Council of Europe, Report by the European Commission for Democracy through Law (Venice Commission) on the relationship between freedom of expression and freedom of religion, CDL-AD(2008)026 (2008).

149 Council of Europe, European Ministerial Conferences on Mass Media Policy and Council of Europe Conferences of Ministers responsible for Media and New Communications Services, https://rm.coe.int/168064 61f b (2013).

150 Council of Europe, Recommendation of the Committee of Ministers to Member States on Combating Hate Speech, CM/Rec(2022)16, https://se arch.coe.int/cm/Pages/result_ details.aspx?ObjectId=0900001680a67955.

151 European Parliament, Proposal to Extend the List of EU Crimes to All Hate Crimes and Hate Speech, COM(2021)0777, https://www.europa rl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/f ile-hate-crimes-and-hate-speech#:~:text=On%209% 20December%202021 %2C%20the,out%20in%20Article%2083(1) (2021).

siderations on the online applicability of the legal framework regulating hate speech.

First, regarding the representation of people targeted by hate speech, it is positive to note the Recommendation clarifies the legal framework for support of those targeted by hate speech. The first accomplishment is that CM/Rec(2022)16 refrains from employing the expression "victims of hate speech". This is a valuable development as there has been a growing advocacy work done in the past decade alerting the use of the word "victim" suggests an inescapable position of fragility, which may result in double victimization of those targeted by hate speech. It is therefore recommended to use the expression "people targeted by hate speech."[152] The second accomplishment is that this Recommendation shares an official disclaimer that, even if race is included as a potential factor in discriminatory practices,[153] this is a highly contested concept. Still, the recommendation could have gone beyond this position by acknowledging race is a concept created during the colonial period by white supremacist colonial powers as a means to justify their colonial project characterized by discriminatory and criminal policies.

CM/Rec(2022)16 also remarkably reflects ongoing debates about the prevention of marginalization practices of minority groups when it requires that MS ensure translations into minority languages,[154] about the cumulative effect of systematic and long-term exposure to hate speech[155] and about the increased harm of hate speech when considering the intersectionality of different systems of oppression for people targeted by hate speech.[156]

Nevertheless, two points could have been addressed differently to better represent those targeted by hate speech. First, despite being understandable that referring to age and gender-sensitive hate speech policies serves illustrative purposes,[157] using "all-inclusive approaches" would have better reflected the intersectionality of groups targeted by hate speech.[158] Similarly, when referring to the queer community, it would have been more inclusive to use queerphobia instead of "LGBTI" (para. 11(d)). LGBTI does not include people who are questioning or who do not identify with other labels but are still within the

152 This research acknowledges this debate and aligns with the advocates of the expression "people targeted by hate speech." However, given the need for an extensive utilization in this research and given the fact that the EU legal system uses "victim" in its EU Victims Directive, for these practical reasons this research will use victims.

153 CM/Rec(2022)16, para. 1(2).

154 CM/Rec(2022)16, Preamble.

155 CM/Rec(2022)16, para. 1(6)(d).

156 CM/Rec(2022)16, para. 58.

157 CM/Rec(2022)16, para. 1(6)(d) and para. 58.

158 The monitoring rounds of the EU Code of Conduct on countering illegal hate speech online show that the main reported types of hate speech in Europe are sexual orientation, xenophobia and hate towards the Roma community. *See* 6th Evaluation of the Code of Conduct, (Oct. 7. 2021), https://ec.europa.eu/info/sites/default/files/factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en_1.pdf, 4.

main targets of hate speech, and "queer" is often used as an umbrella term to refer to all LGBTI people and would have thus been a more inclusive term.[159]

Second, in regard to the conceptualization and regulation of hate speech, this Recommendation provides legal clarity in many aspects. First, although there is no specific human right protecting people from being targeted with hate speech, the Recommendation clarifies in its Preamble that the regulation of hate speech requires a careful balance between the right to private and family life,[160] the prohibition of discrimination[161] and the right to freedom of expression.[162] Notwithstanding, even if the prohibition of abuse of rights[163] is also mentioned in the Preamble, it could have also been mentioned together at the start of the Preamble to better explain the ECtHR legal framework on hate speech. It would have been clearer to explain the legal system altogether, i.e., either the hateful expression is the most serious form of hate speech and therefore an abuse of right under Article 17 ECHR, or it is a violation of the right to private and family life (Article 8 ECHR) and of the prohibition of discrimination (Article 14 ECHR) and therefore should be limited as per the conditions to limit freedom of expression contained in Article 10(2) ECHR. Finally, when referring to Article 17, the Recommendation only mentions that it covers expressions aiming at destroying any rights in the ECHR.[164] But as per the case law of the ECtHR, Article 17 ECHR also covers expressions that aim to limit the rights in the ECHR beyond the limitations allowed for in the ECHR.[165]

Additionally, this Recommendation explains the three types of harmful expressions and clarifies the relation with hate speech.[166] To simplify, it explains that hate speech is a term informally used across many research fields to cover a spectrum of expressions including: (a) criminal offences; (b) problematic expressions, which although not criminal offenses, could be prohibited under civil or administrative law; and (c) content which bears no legal implications but still raises issues of respect and tolerance and should be addressed through culture and education. However, and quite importantly, in legal terms, this Recommendation confirms in paragraph 1(3) that the term "hate speech" should only be used for expressions that are (a) criminalized, and for expres-

159 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.

160 ECHR, Art 8.

161 ECHR, Art. 14.

162 ECHR, Art. 10.

163 ECHR, Art. 17.

164 CM/Rec(2022)16, Preamble para. 12.

165 *See also, Factsheet – Hate Speech*, Press Unit, European Court of Human Rights (June 2022).

166 This explanation aligns with the United Nations' view on the relation between hate speech and harmful speech. *See* Ann. Rep. of the U.N. High Comm'r for Hum. Rts.: addendum, ¶ 12, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013).

sions that, although not amounting to criminal offenses, are (b) prohibited if restricted in line with the conditions laid down in Article 10(2) ECHR.

This distinction is a key contribution from this instrument as it justifies why no legal text should refer to hate speech as illegal or legal. In fact, this means hate speech is always illegal and victims can always seek legal redress, thus there is no need to include "illegal." Still, when referring to applicability of criminal law in the Preamble, it would have added legal clarity had the text mentioned that only the most serious forms of hate speech should be criminalized, and not "some." It is also positive that this Recommendation recognizes the importance of having a comprehensive system composed of various types of measures to address hate speech. It does not only make reference to criminal, civil and administrative law, but it also directs to education and training on human rights, especially as a means to respond to expressions with no legal implication, which still raises issues of tolerance and respect (para. 1(3)(b)).

Additionally, this instrument clarifies the legal framework to assess the severity of the hateful expression as well as the liability framework by referring not only to the conditions for the restriction of the right to freedom of expression as per Article 10(2) ECHR but also to the contextual variables[167] evaluated by the ECtHR on hate speech cases (para. 1(4)). Still, it would have been important to acknowledge the following elements in the description of the contextual variables: the need to investigate historical oppressions as part of the political and social contexts; the need to expressly consider the intersectionality impact of hate speech; and finally, the debate about the element of intent and how some scholars question its relevance for context analysis.[168] Regarding the latter, the element of intentionality was, however, challenged in the dissenting opinion by Judges Ryssdal, Bernhardt, Spielman and Loizou in Jersild v. Denmark when they admitted that good intentions are not enough when provoking racist statements (related to a journalists' intention in the case). This is because, oftentimes, intention is an element that is difficult to assess and also because the harm through the incitement to violence, discrimination or hatred lies precisely in the incitement against democratic and human rights values.

This Recommendation also clarifies the legal threshold for an insult to be considered criminalized hate speech when in paragraph 11(d) it clarifies that racist, xenophobic, sexist and LGBTI-phobic insults are amongst the most serious forms of hate speech and therefore subject to criminal liability. This is a very relevant explanation because it changes the usual doctrine narrative about the conditions for the restriction of freedom under the ECHR. While

167 *See* Section Part 3.2.3.

168 Katharine Gelber (2021) Critical review of International Social and Political Philosophy, 24:4, 402, available at <https://doi.org/10.1080/13698230.2019.1576006> accessed 22 November 2024.

any legal debate about freedom of expression has since the Handyside judgment started by highlighting that freedom of expression encompasses ideas that "offend, shock or disturb,'' it is now clear that in the case of racist, xenophobic sexist and LGBTI-phobic offenses or insults these are to be considered hate speech under criminal law. This threshold follows the legal framework provided for in the APCC and in the Istanbul Convention, thus increasing the legal coherence at the Council of Europe amongst instruments regulating hate speech.

Another achievement in this recommendation is the inclusion of a clear identification of the various stakeholders involved in the regulation of online hate speech, i.e., internet intermediaries, public officials and bodies, media professionals, and civil society organizations. Still, the responsibilities of some stakeholders are vague or described in a way that does not fully reflect the ECtHR case law. For instance, in reference to the responsibilities of public officials, the Recommendation merely suggests that they *avoid* engaging in hate speech (paras. 28 and 29) while the Court's position is that public figures, and especially politicians, have the *duty to refrain* from engagement in hate speech.[169] A good development is that the Recommendation expressly acknowledges that the positive obligations of Member States to prevent human rights violations also apply in the digital environment (Preamble).

Third, regarding the main elements of this Recommendation with implications for the regulation of hate speech in the online environment, it is positive to note that, when discussing the contextual variables to assess the severity of the hateful expression, one of the elements to consider is how the expression is disseminated or amplified. This clear recognition is essential to adequately represent the usually increased damage caused by online hate speech given its elevated reach and fast dissemination.

Turning to content moderation practices, the main consideration is the different threshold between paragraph 16, where the requirement is that Member States remove all hate speech offline or online, and paragraph 32, requiring internet intermediaries to remove only the most serious cases of online hate speech. This can perhaps be understood based on the premise that Member States are obliged to comply with human rights, while the human rights framework is a priori not legally binding upon internet intermediaries. Following this rationale, it is a good legal strategy to require online platforms to remove only the most serious cases of hate speech to avoid over-removal of content, which may not be legally considered hate speech. Nevertheless, it remains unclear how internet intermediaries should moderate the less severe cases of hate speech. In this regard, this Chapter suggests that the CM/Rec(2022)16 could have provided better guidance for the less severe cases by suggesting, for example, that such hate speech should be blocked, labelled

169 *See, e.g.*, Féret v. Belgium, App. No. 15615/07, 573 (July 16, 2009).

and communicated to the respective public oversight body or law enforcement for investigation. To clarify, this Chapter concurs that platforms should remove hate speech actionable under criminal law and suggests platforms should block and label hate speech actionable under civil or administrative law. The CM/Rec(2022)16 could have set this framework more expressly at least for the larger internet intermediaries.

The Recommendation also provides more general guidance on business models and content moderation. It is positive to note the suggestion of the decentralization of content moderation practices because this is a means to better achieve a good contextualization of the expression (para. 33). Additionally, it is noteworthy that even if automation or artificial intelligence systems are deployed, these should still be overseen by human moderation (para. 33) as well as the requirement that human moderators receive training to be up to date with human rights standards (para. 34). Finally, it is remarkable that the Recommendation requires internet intermediaries to *ensure* their business models are not grounded on strategies that directly or indirectly increase hate speech prevalence (para. 36). Including a requirement for businesses to proactively ensure algorithms do not promote hate speech is a completely innovative and needed[170] legal strategy.

With respect to the cooperation between internet intermediaries with law enforcement, the Recommendation in paragraph 22 seems to require that internet intermediaries report only criminal hate speech to law enforcement. Still, if this provision is read in conjunction with paragraph 19 where it is prescribed that all online hate speech be reported to public authorities, it should be concluded that internet intermediaries should report any hate speech to the competent public authorities. This cooperation is all the more important given the requirement in paragraph 16 that Member States and public authorities should remove all hate speech offline and online. Moreover, it is also important to have effective cooperation between internet intermediaries and public bodies to ensure Member States report on hate speech statistics as required in paragraph 25. Notwithstanding, the Recommendation could have clarified the need to have standardized indicators to study hate speech statistics. In fact, the disconnect between the disaggregated data is one of the main current challenges with the reporting as part of the monitoring rounds for the European Code of Conduct on countering illegal hate speech online. This would have presented a good opportunity to provide guidance to European society.

There are also special considerations concerning the responsibilities of internet intermediaries to support those targeted by hate speech. It is positive to note the Recommendation focuses on remedial processes ultimately facili-

170 *See, e.g.*, Jim Waterson & Dan Milmo, *Facebook Whistleblower Frances Haugen Calls for Urgent External Regulation*, Guardian (Oct. 25, 2021),-https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation.

tated through independent judicial reviews (para. 20) and that it requires short and concise explanations to all affected by online hate speech moderation practices (para. 23). However, the remedial responsibilities of internet intermediaries could have been emphasized under section 5 of the Recommendation. For example, involving internet intermediaries in facilitating counter-narratives, or even in cases where the online platform facilitated the dissemination of hate speech they are themselves to be held liable, to provide adequate reparations to those targeted.

In terms of the liability regime for the internet intermediaries, the Recommendation misses the opportunity to address the current preference of some Member States for the regulation of online hate speech through the imposition of fines upon internet intermediaries. This practice can lead to limiting the right to freedom of expression beyond the limits prescribed by ECHR and can result in limiting access to the information and to the media or in worst cases limit the protection of human rights activists.[171] It is positive to see a reference to the importance of considering the various sizes and types of internet intermediaries (para. 21).[172] However, the impact of the different sizes of internet intermediaries for the liability framework could have been more clearly explained. For example, according to the type of platform, different requirements in terms of content moderation practices could be adopted.

2.3.5 Overview Council of Europe, Hate Speech and Historical Oppression

In summary, when analysing the jurisprudence of the Court on Article 10(2) in the light of the original conceptualization by Matsuda[173] of hate speech as expressions perpetuating historical or systematic oppressions, two remarks are in order. First, though it should be recognized that the Court does include political and social background considerations as contextual variables, it fails to consistently refer to historical or systematic oppressions as a key element of hate speech. Similarly, though the Court acknowledges the importance of taking into account the victims' perspectives, it fails to consistently recognize the intersectionality of systems of oppression perpetuated by hate speech.

Regarding other treaties regulating hate speech, these broadly align with the jurisprudence of the ECtHR on the regulation of hate speech and, in some cases, even go beyond to expressly acknowledge the list of protected categories

171 Avi Asher-Schapiro & Ban Barkawi, *'Lost Memories': War crimes evidence threatened by AI moderation*, REUTERS (June 19, 2020), https://www.reuters.com/article/us-global-socialmedia-rights-trfn-idUSKB N23Q2TO.

172 Note the need to take into account the different sizes and functions of internet intermediaries in the design of the liability framework also aligns with the legal framework established by the DSA in the EU.

173 The conceptualization of (racist) hate speech introduced by Matsuda is explained in Section 2 of this Chapter.

as open-ended (e.g. the Istanbul Convention). This recognition of the open-ended conceptualization of the protected groups aligns with the intersectionality theory and with critical legal theory. Still, these treaties fail to expressly mention the element of historical oppression key to the identification of groups targeted by hate speech.

In reviewing the non-treaty initiatives in light of the original conceptualization of hate speech, these initiatives fail to *formally* establish the historical and intersectional elements of oppression in hate speech conceptualized by critical race scholars as part of the contextual variables to be assessed in hate speech cases. Still, progress towards legal coherence has been made and the progressive critical approach adopted in two of the initiatives in mentioning the intersectionality of different systems of oppression, i.e., CM guidelines on upholding equality and protecting against discrimination and hate during times of crisis and CM/Rec(2022)16, should be acknowledged.

2.4 Approaches to Hate Speech by the European Union

2.4.1 General Principles and Primary Sources

The European Union is built on a set of values prescribed in Articles 2 and 3 of the Treaty of the European Union (TEU), which the Member States (MS) resolve to respect and promote. Article 2 of the TEU explains that the foundational values for membership to the EU are human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of minorities. Article 3 clarifies that the EU shall offer its citizens freedom, security and justice.

Although the EU had historically focused on economic integration, it soon acknowledged the growing need to promote human rights.[174] The Court of Justice of the European Union (CJEU) noted in the 1970s that the protection of human rights was a general principle of law for the EU and confirmed that the EU took the ECHR as its main source of inspiration.[175] In Article 6(2) of the TEU, the EU commits to accede to the ECHR, to take the rights in the ECHR as its general principles, and to follow the interpretation of the ECtHR.

The alignment of the EU with the ECHR is also prescribed in the Charter of Fundamental Rights of the EU (CFREU). The CFREU is the EU's leading treaty for the protection of human rights and therefore a key instrument in the regulation of hate speech. The CFREU seeks to ensure European coherence when in Article 52(3) it contains that for articles with corresponding rights

174 Case 29/69 Stauder v. Stadt Ulm, 1969 ECR 419.

175 CJEU, Case *11-70 Internationale Handelsgesellschaft*, paragraph 4; Case *C-60/84 Cinéthèque*, paragraph 26, available at <https://e-justice.europa.eu/563/EN/part_i__protecting_fundamental_rights_within_the_european_union#p1s2.3> accessed 22 November 2024.

in the ECHR, "the meaning and scope of those rights shall be the same"[176] and, likewise, the jurisprudence of the ECtHR over such rights is also applicable to the EU. It should be noted that the EU reserves the right to grant a higher level of protection of human rights than the protection conferred by the ECHR. The complementarity between the EU and the CoE for the protection of human rights has been extensively debated and, in 2005, the CoE suggested the EU should even transpose aspects of other CoE Conventions when within its competence as per the EU Law.[177]

The EU is specifically committed to fighting discrimination based on sex, racial or ethnic origin, religion or belief, disability, age or sexual orientation when designing and implementing its policies (Art. 10 of the Treaty on the Functioning of the European Union (TFEU)). This principle is reinforced in Article 19 of the TFEU, which enables the EU to "take appropriate action to combat discrimination." Additionally, Article 67(3) of the TFEU stipulates that the EU must "ensure a high level of security through measures to prevent and combat crime, racism and xenophobia, and through measures for coordination and cooperation between police and judicial authorities and other competent authorities, as well as through the mutual recognition of judgments in criminal matters." It should be noted that in order to achieve this goal, the EU may adopt measures to approximate criminal laws, namely on hate speech. In fact, the European Commission (EC) communicated its intent to extend the list of the EU crimes to hate speech and hate crime.[178]

Applying this framework to CFREU articles impacting the regulation of hate speech (e.g. Article 1 on human dignity, Article 11 on freedom of expression, and Article 21 on the right to non-discrimination), the CJEU must follow as the minimum standards the conditions for limitation of freedom of expression as per Article 10(2) and adhere to standards developed by the ECtHR in related jurisprudence.[179] For example, due to the wording "such as," the prohibition of discrimination in the CRF (Art. 21 CFR) reflects an open-ended list of impermissible grounds for discrimination, which is also mirroring the formulation in the ECHR (Art. 14). While the CFREU contains an open-ended list of impermissible grounds for discrimination, it also contains specific provisions on certain impermissible grounds such as the respect of diversity (Art. 22), gender equality (Art. 23), age (Arts. 24 and 25), and disabilities (Art. 26).

176 Tarlach McGonagle, *Chapter 24 Free Expression and Internet Intermediaries: The Changing Geometry of European Regulation*, 17 (pre-publication).

177 Council of Europe, Ministers' Deputies CM Documents, Action Plan, CM(2005)80 final 17 May 2005, Appendix I, Guideline 5.

178 Council of Europe, CDADI (n 144).

179 EU Network of Independent Experts on Fundamental Rights, Commentary of the Charter of Fundamental Rights of the European Union (2006) 400.

2.4.2 Secondary Sources

Focusing now on the secondary sources of law of the European Union regulating hate speech, this Section highlights, first, both the Council Framework Decision on combating certain forms and expression of racism and xenophobia by means of criminal law (Framework Decision)[180] and the Audiovisual Media Services Directive (AVMSD). After that, this Section expands on two resolutions adopted by the European Parliament (EP) relevant to explain the regulatory framework on hate speech and to contextualize the ascension of hate speech in the agenda of the EU. This Section also explores three legislative avenues:[181] (1) the EC Regulation of the EP and of the Council on a Single Market for Digital Services (Digital Services Act, DSA);[182] (2) the Regulation of the EP and of the Council Laying down harmonized rules on artificial intelligence (Artificial Intelligence Act, AI Act);[183] and (3) the Directive of the EP and of the Council on combating violence against women and domestic violence. This Section will also explore the intent of the EC communication from December 2021 to extend the list of the EU crimes to hate speech and hate crime.

Similar to the analysis of non-treaty initiatives at the Council of Europe, the study of the responsibilities of private actors when regulating online hate speech (procedural regulation) will not be extensively developed. Although some initiatives on business and human rights will be mentioned for ease of reference on instruments impacting the regulation of online hate speech, this Chapter focuses instead on clarifying the main elements in the conceptualization of hate speech (substantive regulation).

The Framework Decision criminalizes two types of speech – (1) publicly inciting to violence or hatred and (2) publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity and war crimes – when they are directed against a group of persons or a member of such a group defined by reference to "race, colour, religion, descent or national or ethnic origin." From a European human rights perspective, it is problematic that the list of protected grounds is limited to "race, colour, religion, descent or national or ethnic origin." Still, the Framework Decision explicitly suggests that Member States may adopt provisions in national law of crimes "against

180 Acts Adopted Under Title VI of the EU Treaty: 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, 328 OFF. J. OF THE EUR. UNION 55, 55–58 (2008).

181 At the time of publishing, these three legislative acts had not been adopted yet and their analysis had thus been grouped under 'legislative avenues'.

182 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.

183 *Europe: Artificial Intelligence Act Must Protect Free Speech and Privacy*, ARTICLE19, https://www.article19.org/resources/europe-artificial-intelligence-act-must-protect-freedom-of-expression-and-privacy/.

a group of persons defined by other criteria than race, colour, religion, descent or national or ethnic origin, such as social status or political convictions"[184] This wording does not, however, refer to misogynistic or queerphobic speech. In conclusion, the best regulatory strategy to ensure protection of all people targeted by hate speech would have been to leave the list officially open-ended by following Article 21 of the EU CFR. Had the EU legislator chosen an explicit open-ended clause, it would have certainly increased the legal protection of marginalized groups from hate speech. Similarly, though the Framework Decision also calls upon its Member States to guarantee that racist or xenophobic motivation is considered an aggravating factor in criminal law (Art. 4), this approach also does not explicitly mention queerphobia or misogyny motivations.

Another remark is that the Framework Decision lays the emphasis on criminalization, which may wrongfully suggest that other responsive and preventative means should not be prioritized.[185] Similar to the previous analysis on protected groups, even though this instrument does allow for a margin of consideration of severity (Article 2 specifies that Parties may choose to punish only conduct that is likely to disturb public order or which is threatening, abusive or insulting), it would have certainly been a better legislative strategy to include the possibility of various legal (including civil and administrative) and educational measures upfront. Furthermore, the prioritization of criminal law undermines the frequent scepticism of hate speech victims to report it to law enforcement entities for fear of double victimization. To improve the protection of victims within the scope of the Framework Decision, the EC called for the complementary application of the Victims Directive.[186]

The AVMSD governs EU-wide coordination of national legislation on all audiovisual media – traditional TV broadcasts, on-demand services and video-sharing platforms. As per its latest review from 2018, the VMSD recognizes in Recital 45 the problems of increased hate speech and harmful content online. Article 28b states that video-sharing platforms should be required to "take appropriate measures to protect the general public from content that contains incitement to violence or hatred directed against a group or a member of a group on any of the grounds referred to in Article 21 of the CFREU or the dissemination of which constitutes a criminal offense under Union law." As seen, Article 21 of the CFREU postulates a broad understanding of impermis-

184 Acts Adopted Under Title VI of the EU Treaty (n 180), at 56, Preamble, Para. 10.

185 Despite the reference in the Preamble, Para. 6.

186 Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law, COM (2014) 27 final, 9 (Nov. 2018).

sible grounds[187] for discrimination including "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."

Nevertheless, despite prescribing that video-sharing platforms take "appropriate measures" to protect the general public from hate speech on the broad accounts as per Article 21 CFR, the AVMSD only requires Member States to make criminal action available for victims of hate speech as per the Framework Decision, which as just mentioned above applies in racist and xenophobic hate speech cases only (Art. 28b(1)(I)).

The EP adopted two resolutions concerning the protection from hate speech raising attention to the importance of strengthening legal protections from hate speech, especially for certain groups. EP Resolution 2013/2543(RSP) strengthens, inter alia, the need to continue extending protection from discrimination on the basis of non-exhaustive grounds (para. F). EP Resolution 2019/2933(RSP) reinforces the need to adopt an expansive consideration of impermissible grounds for hate speech and underlines the need to address offline and online hate speech (para. 8). This Resolution also calls on Member States to develop mechanisms for monitoring, reporting and investigating online and offline hate speech.

At the EU level, there are three main instruments shaping the regulation of online hate speech.[188] First, the regulation on the DSA, which amends the e-Commerce Directive. The DSA aims to update and harmonize across the EU content moderation practices, due diligence rules, and the framework on the liability of internet intermediaries. Although it reinforces the importance to respect fundamental rights, the DSA does not define what is illegal content. This instrument is what is called a horizontal baseline regime as it focuses on procedural safeguards. For the definition of illegal content in the EU, it is essential to analyse the vertical hard-law *lex specialis*. In the case of hate speech, the *lex specialis* is for the time being the Framework Decision. Addi-

187 This phrasing is inspired by the work of Tarlach McGonagle, *Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas, School of Human Rights Research Series*, 44. The present work will refrain from using "protected categories" and opt instead for "impermissible grounds" because the first can lead the reader to assume that certain segments of the population are inherently disempowered and can thereby contribute to a wrongful stigmatization of such identities, which this research strongly wishes to contest. Hence, in search for a more accurate lexicon, this study chooses to utilize "impermissible grounds" to put the emphasis on the role of law to counter hatred and in an effort to lift any stigmatization of traditionally targeted groups.

188 It is important to clarify that legislative instruments in the form of a regulation are binding on all MS of the EU and do not need to be transposed as they enter into force on domestic legal systems from the moment that they are adopted by the EU legislators. The EU only resorts to regulations in frameworks that are of high relevance to the fulfilment of the EU general principles and when it is politically feasible to require harmonization and standard application throughout the whole EU.

tionally, the DSA places high attention on commitments already agreed in self-regulation practices, such as the Code of Conduct on countering illegal hate speech online.[189] In a complementary manner, the EU AI Act, also addresses online harms by establishing procedural safeguards. Still, both instruments focus on the procedural law rather than the substantive definitions of illegal content, thus not defining hate speech.

Second, in 2024, the European Parliament and the Council adopted the Directive on combating violence against women and domestic violence,[190] which criminalizes cyber stalking; cyber harassment; and cyber incitement to violence or hatred. The new rules in this Proposal strengthen the access to justice for victims and complement the DSA. As this legislative instrument in the form of a Directive, though binding on Member States, these have discretion in transposing it into the domestic legal system. Still, it certainly brings legal cohesion on the regulation of hate speech on the grounds of gender and sex.

Finally, the EC also recently expressed in December 2021 its intent to extend the list of the EU crimes to hate speech and hate crime given its growing impact, especially with technological changes, and given the lack of coherence in the criminalization of hate speech and hate crime among Member States.[191] This initiative aligns with the DSA and the Directive combating violence against women and domestic violence. The collection of these two legislative initiatives would add to the *lex specialis* on hate speech and potentially amend the discussed shortcomings in the Framework Decision.[192] The EC has argued in favour of European standardization stating that hate speech: contains a cross-border dimension; is a relevant area of crime particularly serious as it undermines EU fundamental rights in Art. 2 and 6 of the TEU and the CFR; presents recent worrying developments, especially with online hate speech. The Council of the EU and the EP now have to agree that these areas represent another area of crime in need of standardization across the EU.

189 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.

190 European Union, Directive (EU) 2024/1385 of the European Parliament and the Council of 14 May 2024 on combating violence against women and domestic violence, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401385> accessed 20 November 2024; *see also* European Commission Press Release IP/22/1533, International Women's Day 2022: Commission Proposes EU-Wide Rules to Combat Violence Against Women and Domestic Violence (Mar. 8, 2022).

191 Communication from the Commission to the European Parliament and the Council, COM (2021) 777 final (Sept. 12, 2021); *see also* European Commission Press Release IP/21/6561, The Commission proposes to extend the list of 'EU Crimes' to hate speech and hate crime. (Dec. 9, 2021).

192 *Id.*

2.4.3 Soft Law

The EU Code of Conduct on countering illegal hate speech online (Code) was adopted in 2016 and is a non-binding self-regulatory instrument.[193] The Code does not define hate speech. Instead, it adopts the definition of "illegal hate speech" of the Framework Decision, which only grants protection on the grounds of race, colour, religion, descent or national ethnic origin. There are two problematic aspects with this framework.

First, the distinction between legal and illegal hate speech is problematic as the utilization of *illegal* hate speech seems to imply there is legal hate speech. It is true the term hate speech has been *informally* used to refer to expressions regulated as criminal offenses, civil or administrative breaches, or even to harmful expressions not comporting legal consequences. Nevertheless, the ECtHR has only referred to hate speech to cover criminal offenses or civil or administrative breaches. Therefore, the third category of expressions, i.e., harmful expressions without legal implications, should not be considered hate speech at all. The utilization of legal hate speech for this third type of expressions induces legal unclarity and enforceability which compromises the regulatory efforts of regulating hate speech.

Second, the restriction of the impermissible grounds of hate speech to accounts of race or xenophobia is not aligned with the case law of the ECtHR and with the open-ended list of grounds protected from discrimination under Article 21 of the CRFEU.

Some remarks are necessary with regard to the procedural requirements stemming from the Code. It is confusing to note that despite pointing in the direction of a definition of hate speech, the Code prescribes that the IT companies will nevertheless review the illegality of the content against their community guidelines and not against the Framework Decision or national laws transposing it. Additionally, it is concerning that the Code gives the lead in this balancing of rights to the IT Companies without instructing them on the broader European human rights law and interpretative case law. The latter remark is particularly important in the case of IT companies of substantive size and thereby potentially delivering services comparable to public service. Moreover, the Code does not require reporting on the methodology for monitoring practices of hate speech which renders any study on the efficacy of the measures impossible and therefore does not comply with the conditions for restriction of freedom of expression as per Article 10(2), i.e. in that it is important to prove that the restriction is prescribed by law, in pursuit of legitimate interests and necessary in a democratic society.

---

193 *European Union Code of Conduct on Countering Illegal Hate Speech Online*, EUROPEAN COMM., https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discriminati on/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech -online_en.

2.4.4 Overview European Union, Hate Speech and Historical Oppressions

In summary, the legal strategies of the EU to conceptualize hate speech fall short from formally aligning with the original conceptualization by Matsuda[194] of hate speech as expressions perpetuating historical, systematic and intersectional systems of oppression.

The EU has in different moments adopted conflicting positions in regards to the intersectionality of systems of oppression. Although the AVMSD refers to the open-ended list of impermissible grounds for hate speech as per Article 21 of the CFR,[195] in the Framework Decision the EU limited the protection to people targeted on the basis of race, colour, religion, descent or national ethnic origin. This shows internal legal incoherence and a significant deviation from the intersectionality theory, which highlights the need to grant protection to any person from hateful expressions reinforcing a historically vertical relationship (among others, women, the queer community, and persons with disabilities).

2.5 Main Elements of Hate Speech in the European Context

The following Section describes the main legal elements of hate speech in the European context by building on the findings from its conceptualization at the Council of Europe and at the European Union. This exercise will consider in particular how such main elements of hate speech in the European context relate to the legal theoretical foundations on hate speech stemming from critical legal theory, specifically from critical race theory. It is important to reiterate that this concluding Section in the Article does not purport to reach a clear definition of what hate speech is. Instead, the main goal is to critically distil the main elements surfacing in the European context about the regulation of hate speech.

2.5.1 What Is Hate Speech and What Does It Do?

Although there is no agreed, legally binding definition of hate speech at the European level,[196] the findings from the analysis of the instruments at the Council of Europe and at the European Union level help present what the

194 The conceptualization of (racist) hate speech introduced by Matsuda in Section 2 of this Chapter.

195 European Union, *Charter of Fundamental Rights of the European Union*, 26 October 2012, 2012/C 326/02, https://www.refworld.org/docid/3ae6b3b70.html .

196 Neither at the European nor at the international level.

general common understanding of hate speech is per the European human rights standards.

Hate speech can be broadly conceptualized as any kind of expression that incites, promotes, spreads or justifies violence, hatred or discrimination against one person or a group of people based on presumed or real identity characteristics such as race, religion, sex, gender, sexual orientation, citizenship, national or ethnic origin, age, or disability. Four explanatory remarks are due.

First, the meaning of "expressions" should be construed in a broad manner and be understood as *any* expression be it verbal, non-verbal (such as facial expressions, gestures, pictures, signals), shared in person or online (including for example GIFs and memes). This is supported by CM/Rec(2022)16 when it prescribes that hate speech includes "all types of expressions." Furthermore, this is also the understanding in the Istanbul Convention when in Article 40 it includes "verbal and non-verbal expression."

Second, the impact of the expression should also be construed in an expansive way to include dissemination, promotion, and support. This is supported by CM/Rec(2022)16 when in paragraph 1(2) it refers to expressions that "incite, promote, spread or justify violence, hatred or discrimination."[197] Moreover, this is also aligned with the inclusion of dissemination as a qualifying act of hate speech by APCC Article 3 and ECRI GPR 6 as well as with the inclusion of aiding and abetting contained in the Istanbul Convention Article 41.

Third, the list of identity characteristics protected from hate speech is to be broadly conceptualized, too, and to be considered open-ended. The ECtHR, to date, in its interpretation of the ECHR has ruled on cases of racial, ethnic, religious, and homophobic hate speech. In its case law, there is no indication of limits regarding the impermissible grounds of hate speech. Furthermore, taking as the point of departure human rights provisions on non-discrimination, both Article 14 ECHR and Article 21 of the CFREU include "such as" before illustrating the list of impermissible grounds for discrimination. However, some confusion in the application of such a standard may arise when looking into some sector specific regulatory instruments on the criminalization of hate speech both at the EU and at the CoE level.

For example, within the CoE regulatory framework, there are distinctions in the protection granted from hate speech by sector specific treaties. For example, while the APCC only protects against racism and xenophobic speech, the case of the Istanbul Convention is more complex. Though the latter stresses the need to protect people from hate speech on the basis of their sex and gender, the drafters of the Explanatory report suggested the adoption of an open-ended list for grounds for non-discrimination, including gender, sexual orientation, gender identity, age, state of health, disability, marital status, and

197 U.N. Doc. A/HRC/22/17/Add.4 (n 70).

migrant or refugee status.[198] As a result, with this expansive interpretation of the groups protected under the Istanbul Convention from hate speech, the CoE currently grants redress through criminal law to an open-ended list of impermissible grounds.

Within the EU legislative framework, there seems to be a disconnect between the protected categories in the Framework Decision and in the Code of Conduct (which aligns with the Framework Decision) when compared with the conceptualization of protected categories under the AVMSD. To clarify, on one hand, the AVMSD aligns with the CFREU and addresses hateful expressions on the grounds of a wide range of categories including sexual orientation, disability, and age. On the other hand, the Framework Decision and the Code of Conduct address only racist and xenophobic hate speech. Still, it should be highlighted that, for the purposes of criminalization, the AVMSD refers to the hate speech conceptualization of the Framework Decision. As a result, the EU currently only grants access to criminal law to people targeted with hate speech on the grounds of their race, colour, religion, descent, or national or ethnic origin.

As observed by Alkiviadou more generally and emphasized here in the context of the EU, the European legal strategies can be interpreted as validating a "hierarchy of hate,"[199] where only certain forms of hatred are classified as impermissible and only certain affected groups could seek redress under criminal law. This is a point of concern, particularly in the EU system, because it overlooks research on enhanced prevalence of hatred based on victims' intersectional characteristics and because recent data shows that online hate speech across the EU level is mostly registered toward group's identities on the basis of sexual orientation.[200] Furthermore, these legal strategies separating systems of oppression fail to address the works from critical legal and race theorists when they were alerted to the intersectionality of socio-historical contexts of oppression (i.e., how someone targeted by different systems of oppression would need better protection from hate speech; an example of someone targeted by different systems of oppression could be someone racialized, a woman, a queer individual, or a disabled individual). As a result, the only route to progress toward a more coherent and cohesive European human rights framework is to expressly adopt an open-list for grounds prohibiting discrimination, whereby intersectionality of systems of oppression is clearly addressed as a criterion in the legal assessment.[201]

198 Convention on Preventing and Combating Violence Against Women and Domestic Violence (n 130), ¶ 53.

199 Natalie Alkiviadou, *The Legal Regulation of Hate Speech: The International and European Frameworks*, 55 Politièka Misao 203, 223 (2018).

200 Didier Reynders, Directorate-General for Justice and Consumers, European Commission, 5th valuation of the Code of Conduct on Countering Illegal Hate Speech Online (June 4, 2020).

201 Though important to continue to guard against the dilution of grounds.

The fourth and final remark on what hate speech is and what impact it has on its targets relates precisely with the fact that hate speech, as per its foundational conception, targets people who exist in systems of oppression. It is worth reminding that, the original conception in critical legal theory presented by Matsuda, "the hateful message is directed against a historically oppressed group and reinforces a historically vertical relationship."[202] This is a relevant reminder, especially with the current interpretation that the list of impermissible grounds for hate speech is open-ended. The interplay between these two conceptual elements should be translated to mean that, although any person can be targeted by hate speech based on a given group categorization, hate speech primarily functions as a means to keep the targeted group oppressed. A caveat should nevertheless be introduced to grant protection from hate speech in case of a "pressing social need" to protect 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.

With regard to the legal framework for measuring the existence of systems of historical or systematic oppression, this Chapter suggests focusing on analysing the group's representation in political, economic, and social spheres. For instance, if the group has effectively exercises their rights to education, to work, or to access information. In contexts where policies of affirmative action have been designed and implemented, these represent a good framework for examining whether the group endures systematic oppression.

For clarity, this Chapter gives two examples of the application of this framework. For instance, white, heteronormative, cisgender, neurotypical, and able men, in principle belonging to various privileged societal groups, would not be immediately granted protection under hate speech laws unless proven that, with the experience of being targets of such speech, the protection of the group from such speech had become a pressing social need as per the ECtHR case law. This reasoning can be applied to the real case scenario mentioned in the introduction Section of this Chapter. To recapitulate, in the wake of a terrorist attack in London in 2017, a U.S. congressman wrote a Facebook post in which he directly called for the slaughter of all "radicalized" Muslims. This post pleading for imminent violence went untouched by Facebook. Later that year, a post on Facebook by a Black Lives Matter activist saying "All white people are racist" drew a different response. This post was removed, and the account was disabled for seven days. Facebook justified the different responses by stating "radicalized Muslims" was a sub-group of protected groups, while "all white people" was a more encompassing group and therefore deemed it more critical to protect. To properly address such a case, the current European human rights standards would need to have an explicit reference to the

202 Matsuda et al. (n 20), 35.

balancing between systems of oppression in order to deliver on values of democracy and human rights, including the rights of minorities.

Another example is the high levels of hate speech directed at health professionals during the COVID-19 pandemic. Here, although there is not an historical oppression of people belonging to such a group, the high levels of systematic threats had a substantial impact on how such professionals could live their lives and, in this case, they would certainly be protected under hate speech laws.

2.5.2 How to Substantively Regulate Hate Speech

The below Section introduces five main axes or principles for the regulation of hate speech: (i) hate speech is always illegal speech; (ii) hate speech laws aim to protect first and foremost historically oppressed groups; (iii) hate speech is only criminalized in its most serious forms; (iv) if not criminalized, then balancing of rights ideally on a case-by-case basis; and (v) regulating hate speech cannot simply be case-by-case.

2.5.2.1 Hate Speech Is Always Criminal or Illegal Speech

Hate speech should only be used to refer to either criminal offenses or problematic expressions, which although not criminal offenses should be prohibited under Article 10(2) of the ECHR and be legally actionable under civil or administrative law. Applying this framework to the ECHR, the term hate speech should only be used for expressions comprising an abuse of rights (Art. 17 ECHR) and for expressions, which despite being within the remit of freedom of expression (Art. 10(2) ECHR) can or should be legally limited in line with conditions explained in Article 10(2). The legal framework should not use the expression illegal hate speech as it induces the reader into thinking there is legal hate speech, therefore creating challenges of legal clarification and foreseeability. Expressions of intolerance and disrespect not amounting to hate speech but still raising issues of respect and tolerance should be addressed through culture and education.

2.5.2.2 People Targeted By Hate Speech Have Been Historically or Systematically Oppressed

The current European human rights standards need to explicitly acknowledge the linkage between hate speech and the perpetuation of systems of oppression. This is a key acknowledgement to promote the values of democracy and human rights, including the rights of minorities, guiding the European human rights framework. In considering the social background, the ECtHR should

expressly investigate socio-historical systems of oppression.[203] This also aligns with the "positive obligations doctrine" whereby States have a positive obligation to investigate bias indicators (i.e., objective facts or circumstances by which probable motives can be discerned).[204] The necessity test must also consider the victims' perspectives. In looking into the victims' perspective, the relevance of the intersectionality framework must be highlighted to better reflect the "living instrument doctrine," i.e., to have interpretations that adequately reflect present-day realities of targeted groups. The list of impermissible grounds for hateful expressions must be open-ended to account for present day realities of oppression (e.g. queerphobic speech, hate targeting refugees).

2.5.2.3 Hate Speech Should Only Be Criminalized In Its Most Serious Forms

European human rights standards require that only the most serious forms of hate speech be criminalized. The ECtHR adopts an assessment on a case-by-case basis. Nevertheless, its case law under Article 17 should guide the European framework regulating hate speech amounting to criminal offences. The jurisprudence of the ECtHR regarding expressions that constitute the most serious cases of hate speech was summarized in Paragraph 11 of the Appendix to CM/Rec(2022)16 and it includes:

a. public incitement to commit genocide, violence or discrimination;
b. racist, xenophobic, sexist and LGBTI-phobic threats;
c. racist xenophobic, sexist and LGBTI-phobic public insults under conditions such as those set out specifically for online insults in the Additional Protocol to the convention on Cybercrime concerning the criminalization of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189);
d. public denial, trivialization and condoning of genocide, crimes against humanity or war crimes; and,
e. intentional dissemination of material that contains such expressions of hate speech (listed in a-e above) including ideas based on racial superiority or hatred.[205]

Notwithstanding this useful summary of the most serious cases of hate speech contained in CM/Rec(2022)16, the list of impermissible grounds for hate speech should not be strictly interpreted but rather open-ended. That is because it is relevant to investigate the history and intersectionality of systems of oppression in the given context where the hateful expression was communi-

203 Sahana Udupa, *Decoloniality and Extreme Speech*, Paper Presented at the 65th e-Seminar, Media Anthropology Network, European Association of Social Anthropologists 24 (2020).
204 *See, e.g.*, Ibentoba v. Georgia, App. No. 73235/12 (May 12, 2015), https://hudoc.echr.coe.int/fre?i=001-154400.
205 U.N. Doc. A/HRC/22/17/Add.4 (n 70).

cated (as just explained in Section 2.5.2.2). The victims' perspective must be construed to allow for an interpretation that is reflective of present-day realities (e.g. queerphobic speech, hate targeting refugees).

The limited classification of criminal hate speech only for the most serious forms of hate speech enables the consequential application of the right to an effective remedy as prescribed in Article 13 of the ECHR. All other forms of hate speech, illegal but not criminal, can lead to administrative or civil action, thus avoiding legal uncertainty on criminal action and thereby contributing to a more effective right to remedy.

2.5.2.4 If Not Criminalized, Then Need to Balance Human Rights

"Freedom of expression is the condition *sine qua non* for a genuine pluralist democracy":[206] The ECtHR ruled in Handyside v. UK that expressions that "offend, shock or disturb" are generally protected by the ECHR, considering that "such are the demands of pluralism, tolerance, and broadmindedness without which there is no democratic society."[207] States have to justify even minor interferences with Article 10.[208] Expressions that are prohibited but not criminal (thus not triggering the application of Article 17 ECHR) must respect conditions as per Article 10(2) ECHR: provided by law; pursuing one of the mentioned legitimate purposes; and, be necessary in a democratic and pluralistic society, following the ECtHR jurisprudence. The necessity test must encompass the analysis of the following contextual factors: political and social background; intent of the speaker; speaker's status or role in society; content of the expression; extent of the expression; and the nature of the audience. In this exercise of assessing whether a limitation on the right to freedom of expression is necessary in a democratic society, it is important to explicitly account for socio-historical records of oppression and the victims' potential intersectional position between various oppressive systems.

2.5.2.5 Challenges With the Regulation of Online Hate Speech

New technologies, social media and algorithms designed to promote content virality have greatly contributed to the increase of online hate speech. Given the prevalence of hate speech online, it is impossible to assess posts on a case-by-case basis. Regulation of online harms in general, and of hate speech in

206 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*, Seminar on Human Rights for European Judicial Trainers (Sept. 3, 2013). Tulkens defends that freedom of expression is both a safeguard against interference by the State (subjective right) and an objective right, and a means for the establishment of a democratic society.

207 Handyside v. UK (n 102).

208 *See, e.g.*, Ottan v France, App. No. 41841/12 (July 19, 2018), https://hudoc.echr.coe.int/fre?i=001-182627.

particular, focuses on instructing internet intermediaries on minimum human rights procedural safeguards and public monitoring and auditing of the digital platforms' compliance with due diligence frameworks. Internet intermediaries often moderate online content with the use of automated decision-making tools. Internet intermediaries should design and employ automated-decision making tools to identify, ex ante, at least the most serious cases of hate speech, as elaborated in the jurisprudence of the European Court of Human Rights on Article 17 of the European Convention on Human Rights.

2.6 Conclusion

Scholars, practitioners, and policy-makers have long focused on clarifying the definition and status of hate speech in international and regional human rights. However, this has proven to be a challenging process and the absence of a legally-binding definition of hate speech in human rights law has had severe negative individual and societal implications. In an era of digital communication where there is an increased prevalence and reach of hate speech, it is imperative to advance a standardized legal conceptualization of hate speech that is suitable to protect and to present legal remedies for people targeted by such hateful expressions.

This Chapter clarifies the original conceptualization of hate speech advocated by critical race scholars, grounded on the perpetuation of intersectional, historical or systematic oppression. This Chapter then analyses, in the light of that original conceptualization of hate speech, a selection of legal initiatives in the European context, covering treaties and non-treaty initiatives suggested as the most relevant in the regulation of hate speech at the European level. The main treaty instruments analysed are the European Convention on Human Rights and the Charter of Fundamental Rights of the EU. The most relevant EU legal instruments are the Framework Decision on Combating Racism and Xenophobia, the Audiovisual Media Services Directive and the Code of Conduct on countering illegal hate speech online. The main non-treaty instrument is the Recommendation of the Committee of Ministers of the Council of Europe on a comprehensive approach to hate speech (CM/Rec/(2022)16).

The analysis in this Chapter explains the interplay between European regulatory instruments and claims that a more standardized conceptualization of hate speech rooted in the intersectional, historical or systematic systems of oppression perpetuated by hate speech can help reconcile the regulation of hate speech in Europe. If the European regulatory framework to counter hate speech is to uphold values of equality, dignity, pluralism, and democracy, then the most effective and legally coherent manner to achieve that objective is to emphasize the seminal hate speech elements presented by critical race and legal scholars and to emphasize the need to investigate intersectional, historical or systematic systems of oppression perpetuated by hate speech.

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.

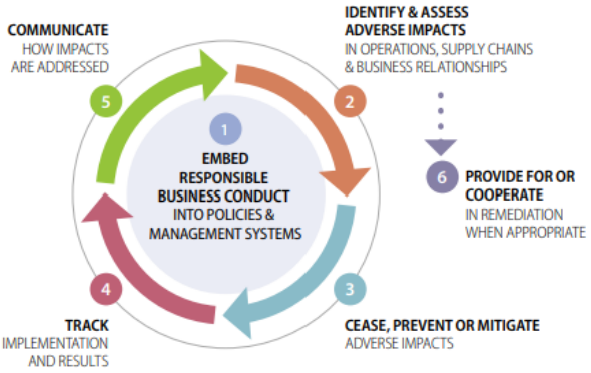

Figure 2 – OECD Due diligence process and supporting measures[82]

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.

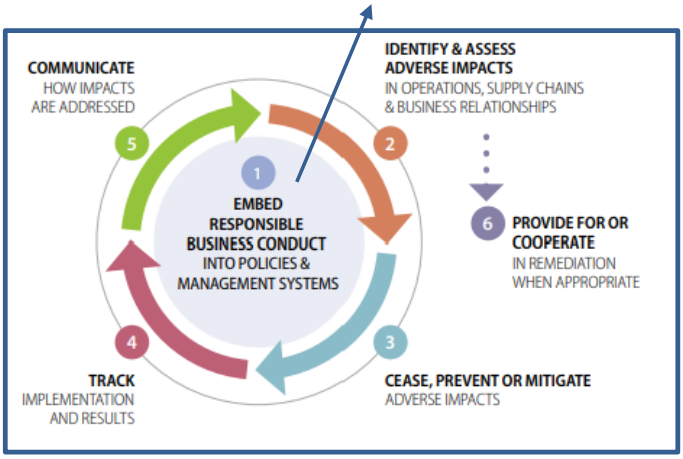

Figure 3 – OECD Due diligence process, including precautionary measures to counter criminal hate speech

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.

4 Human rights responsibilities of online platforms to mitigate criminal hate speech

Disrupting incitement to violence in large groups on end-to-end encrypted services in Europe[12]

ABSTRACT

Business models adopted by online platforms have enabled the proliferation of online hate speech. End-to-end encrypted (E2EE) services have been under increased scrutiny for hosting hate mongers. Legal practitioners and law enforcement struggle to conceptualise the responsibilities of E2EE services to not host hate speech without disproportionately affecting the users' rights to freedom of expression, association, privacy, or data protection. This interdisciplinary study proposes a new legal minimum standard expanding corporate human rights responsibilities of E2EE services to counter a category of criminal hate speech – incitement to violence. We explore the regulation and application of metadata, hashing, and homomorphic encryption to disrupt incitement to violence in large groups on E2EE services.

1 This Chapter was originally published in the Technology and Regulation journal, 2024 (2024): 115-131, in co-authorship with Stephan Raaijmakers and Thijs Veugen. Stephan Raaijmakers is a Senior scientist at Netherlands Organisation for Applied Scientific Research (TNO); Professor of Communicative Artificial Intelligence, Leiden University Centre for Linguistics. Thijs Veugen is a Senior Scientist at Netherlands Organisation for Applied Scientific Research (TNO); Professor of Applied Cryptography, University of Twente.

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; the European Union Directive of the European Parliament and of the Council on corporate sustainability due diligence (CSDDD); and the European Commission 2024 Proposal for a Directive of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse (CSAM Directive). Cross-references should be read as referring to other references within the present Chapter.

4.1 INTRODUCTION

Management boards of online platforms have adopted business models enabling the proliferation of online hate speech.[3] While online hate speech initially appeared on openly accessible platforms,[4] hate mongers are increasingly operating on encrypted services, as these provide higher protection of anonymity, privacy, and thus less accountability.[5] In particular, end-to-end encrypted (E2EE) services have been under higher scrutiny for hosting and facilitating the growth of hate speech.[6]

The migration of hate mongers to E2EE services represents one of the newest regulatory and law enforcement challenges when countering online hate speech, as internet intermediaries[7] and civil society claim that it is technically impossible to detect illegal content in E2EE services without compromising the privacy features.[8] For example, Facebook Help Center states "This means that nobody else can see or listen to what's sent or said – not even Meta. We couldn't even if we wanted to."[9]

How to prevent the proliferation of hate speech on E2EE services? On the one hand, it requires a cautious assessment of the relationship between the right users' human rights and the internet intermediaries' corporate human

3 *E.g.*, Alex Cranz and Russell Brandom, 'Facebook encourages hate speech for profit, says whistleblower' (The Verge, 2021), available at <https://www.theverge.com/2021/10/3/22707860/facebook-whistleblower-leaked-documents-files-regulation> accessed 28 Aug 2023; Karen Hao, 'The Facebook whistleblower says its algorithms are dangerous. Here's why.' (MIT Technology Review, 2021), available at <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/> accessed 28 Aug 2023; Newley Purnell and Jeff Horwitz, 'Facebook Services Are Used to Spread Religious Hatred in India, Internal Documents Show' (The Wall Street Journal, 2021), available at <https://www.wsj.com/articles/facebook-services-are-used-to-spread-religious-hatred-in-india-internal-documents-show-11635016354?mod=article_inline> accessed 28 Aug 2023,.

4 *E.g.*, Noah Giansiracusa, 'Facebook Uses Deceptive Match to Hide its Hate Speech Problem' (Wired, 2021), available at <https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/> accessed 17 October 2023.

5 Tech against terrorism, 'Terrorism use of E2EE: State of Play, Misconceptions, and Mitigation Strategies Report' (2021), available at <https://www.techagainstterrorism.org/wp-content/uploads/2021/09/TAT-Terrorist-use-of-E2EE-and-mitigation-strategies-report-.pdf>accessed 28 Aug 2023, 42-56.

6 Tech against terrorism (n 5).

7 'Internet intermediaries' includes hosting intermediaries, domain providers, search engines, messaging providers, access providers, etc. 'Internet intermediaries' is used interchangeably with 'online platforms', 'AI businesses', or with 'IT companies', depending on the legal instrument under analysis. 'Businesses' and 'companies' are used synonymously. 'Internet intermediaries' includes platforms providing E2EE services.

8 Maria Koomen, 'The Encryption Debate in the European Union: 2021 Update' (Carnegie Endowment for International Peace, 2021), available at <https://carnegieendowment.org/2021/03/31/encryption-debate-in-european-union-2021-update-pub-84217> accessed 28 Aug 2023.

9 Facebook Help Centre, available at <https://www.facebook.com/help/messenger-app/786613221989782?cms_id=786613221989782> accessed 28 Aug 2023.

rights due diligence (HRDD)[10] responsibility to counter cybercrime, in particular criminal hate speech. Thus far, the regulation of HRDD of E2EE services has focused on the prevention of child sexual abuse material. These regulations have been criticized for violating data protection law.[11] On the other hand, given the privacy-preserving features of E2EE, law enforcement bodies lose the typical oversight capacity that they would otherwise have offline. To date, law enforcement techniques in E2EE services have focused on infiltration of groups which has been criticized for violating human rights.[12]

This Chapter addresses this combined legal and technical challenge by focusing on the following research questions: Can there be an innovative and proportional legal interpretation of technological developments that clarifies and expands the HRDD of E2EE services in the European context to not host criminal hate speech in the form of incitement to violence? If so, can this innovative interpretation result in new corporate HRDD responsibility standards for cooperation with law enforcement?

This Chapter provides an interdisciplinary human rights doctrinal analysis of new digital technologies. This research has a European focus, combining analysis of instruments at the levels of the Council of Europe (CoE) and the European Union (EU), given the overall alignment of these two legal regimes.[13] Nevertheless, as the European HRDD framework derives significantly from international standards, there will be occasional reference to international instruments.

Section 4.2. explains the conceptualization of criminal hate speech by critically analysing the European human rights conceptualization in Recom-

10 Both HRDD and internet intermediary liability regimes prevent and address the negative impact of businesses on human rights. However, HRDD and the liability regime differ, as exemplified in the DSA where there are allocated to separated chapters. These regimes are nevertheless related in that liability may follow from non-compliance with HRDD responsibilities.

11 Sabine K. Witting and Gianclaudio Malgieri, "Voluntary detection order under the proposed EU Child Sexual Abuse Regulation violate EU (privacy) law" (European Law Blog, 2023), available at <https://europeanlawblog.eu/2023/05/15/voluntary-detection-orders-under-the-proposed-eu-child-sexual-abuse-regulation-violate-eu-privacy-law/> accessed 28 Aug 2023; "it discourages companies from making their services more secure by developing and deploying encryption.", available at <https://www.bitsoffreedom.nl/2022/05/11/european-commission-wants-to-eliminate-online-confidentiality/> accessed 28 August 2024.

12 EDRi 'How Europol's reform enables 'NSA-style 'surveillance operations'(2021) available at<https://edri.org/our-work/how-europols-reform-enables-nsa-style-surveillance-operations/> accessed 17 October 2023. For a more general study on human rights concerns of law enforcement infiltration, see Katie Pentney, 'Licensed to kill… discourse? Agents provocateurs and a purposive right to freedom of expression' (Netherlands Quarterly of Human Rights, 2021), Vol. 39(3) 241-27, available at <https://journals.sagepub.com/doi/pdf/10.1177/09240519211033429> accessed 28 Aug 2023.

13 Article 52(3) of the Charter of Fundamental Rights of the EU (CFREU) 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.

mendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech.[14] This Recommendation distils the main categories of criminal hate speech found in treaty law[15] and in case law of the European Court of Human Rights (ECtHR).[16] The most relevant category of criminal hate speech for this Chapter is incitement to violence. This section then presents an analysis of the implications of criminal hate speech on E2EE settings. Finally, Section 4.2. clarifies the key human rights safeguards in countering criminal hate speech on E2EE services, such as the operationalization of the legal requirements for restricting freedom of expression, association, privacy, and data protection.

Section 4.3. explains the corporate HRDD responsibilities to counter criminal hate speech in E2EE services. After establishing the general HRDD framework for Artificial Intelligence (AI) businesses,[17] this section applies the HRDD regime to internet intermediaries countering criminal hate speech. The general HRDD instruments analysed are the United Nation Guiding Principles on Business and Human Rights (UNGPs),[18] the EU Corporate Sustainability Due Diligence Directive (CSDDD),[19] and the EU Artificial Intelligence Act (AI Act).[20] The instruments regulating the HRDD responsibil-

14 Council of Europe Committee of Ministers, Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech, available at <https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a67955> accessed 7 Sep 2023. Hereinafter 'CM/Rec(2022)16' or 'the Recommendation'.

15 Such as the European Convention on Human Rights (ECHR) and the First Additional Protocol to the Convention on Cybercrime.

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

17 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. 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.

18 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.

19 European Commission (2022) Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937. The Council of the EU and the European Parliament reached a provisional agreement in December 2023. Janos Allenbach-Ammann (2023) EU Parliament and member states reach deal on corporate due diligence law, *EURACTIV*, available at <https://www.euractiv.com/section/economy-jobs/news/eu-parliament-and-member-states-reach-deal-on-corporate-due-diligence-law/> accessed 5 Feb 2024.

20 European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts COM(2021)206 final. The AI Act was agreed by EU policymakers in December 2023 and approved by the Council of the EU in January 2024. The AI Act enters into force 20 days after publication in the official journal.

ities of internet intermediaries to counter hate speech are: two binding instruments with horizontal application regardless of the type of online content, i.e. the EU Regulation on a Single Market for Digital Services (DSA),[21] and the EU Audiovisual Media Services Directive (AVMSD)[22]; and two sector specific instruments applicable to online hate speech, one of which one is a co-regulatory initiative and another a policy-setting instrument, i.e. respectively the EU Code of conduct on countering illegal hate speech online[23] and the CM/Rec(2022)16. This section then problematises the regulation of E2EE services regarding two alternative types of illegal content by reviewing two instruments: the European Commission (EC) proposed Regulation laying down rules to prevent and combat child sexual abuse (CSAR)[24] and the Regulation to address the dissemination of terrorist content online (TCOR).[25]

Luca Bertuzzi (2024) EU countries give crucial nod to first-of-a-kind Artificial Intelligence law, *EURACTIV*, available at <https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/>accessed 5 Feb 2024.

21 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.

22 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.

23 European Commission (2016) The Code of Conduct on countering illegal hate speech online. The use of 'illegal hate speech' can mislead the reader to consider that there is legal hate speech, which is not accurate. Hate speech is always illegal under civil or administrative law and, in its most severe forms, it can be criminally actionable. For legal coherence, this research refrains from using 'illegal hate speech' unless referring to the title of an instrument.

24 It should be noted that the European Commission published, in 2024, a new Proposal for a Directive to prevent and combat CSAM, *i.e.* European Commission (2024) COM (2024) 60 final 2024/0035 (COD): Proposal for a Directive of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse and replacing Council Framework Decision 2004/68/JHA (recast) available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2024%3A60%3AFIN> accessed 20 November 2024. Notwithstanding, for illustrative purposes of the potential legal challenges, this thesis presents the analysis as originally published in the academic journal, which is based on the 2022 proposal by the European Commission for a Regulation to prevent and combat CSAM, *i.e.* European Commission (2022) COM (2022) 209: Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, available at <https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX:52022PC0209> accessed 7 Sep 2023.

25 European Union (2021) Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 Apr 2021 on addressing the dissemination of terrorist content online, L 172/79, available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0784&qid=1694075338473> accessed 7 Sep 2023.

Section 4.4. delves deeper into the digital technologies and encryption features used for content moderation[26] in E2EE services. This section focuses on metadata, hashing, combined with homomorphic encryption.

Section 4.5. proposes a new legal HRDD standard expanding corporate HRDD of E2EE services and clarifying their framework for cooperation with law enforcement bodies in the context of incitement to violence in large group chats. We analyse the application of the HRDD regime coupled with homomorphic encryption, metadata, and hashing to selected criminal hate speech inciting to violence.

4.2 Criminal hate speech as cybercrime

4.2.1 Incitement to violence as criminal hate speech

Currently, there is no legally binding definition of hate speech in international or European human rights law. Nevertheless, it is possible to find the main elements of hate speech in Recommendation CM/Rec(2022)16 on combating hate speech. Though not legally binding, this recommendation adopted by the statutory decision-making body of the CoE clarifies the states' obligations and businesses' responsibilities based on existing human rights standards deriving from treaty law, ECtHR jurisprudence,[27] and other standard-setting instruments.

CM/Rec(2022)16 explains that, from a legal perspective, hate speech can be subdivided into two categories: (1) the most serious cases of hate speech which should be criminally actionable and, (2) hate speech prohibited under civil or administrative law.[28] Outside the legal framework, the term hate speech is also wrongly used to refer to a third type of speech, *i.e.* harmful expressions, which are not severe enough to be prohibited under the ECHR.[29]

This Chapter focuses on category (1), i.e. criminal hate speech, because there is a clearer understanding at the European level of its main elements. This understanding offers a more precise common ground under which specific HRDD responsibilities can be required of internet intermediaries. The emphasis on criminal hate speech is all the more important since the European Commis-

26 Though referred to as "content moderation techniques", this research acknowledges these techniques could also be referred to as content detection techniques.

27 *E.g.*, for a good summary of ECtHR case law see ECtHR (January 2023) Factsheet – Hate Speech, available at <https://www.echr.coe.int/documents/d/echr/FS_Hate_speech_ENG> accessed 7 Sep 2023.

28 CM/Rec(2022)16, Appendix, Para. 3.

29 Human rights standards suggest that these harmful but lawful expressions should be countered with alternative responses to legal action, such as education, dialogue, and awareness-raising activities. CM/Rec(2022)16, Appendix, Para. 31 and 56.

sion communication about its intention to extend the list of EU crimes to hate speech.[30]

CM/Rec(2022)16 presents a summary of the main categories of criminal hate speech.[31] This conceptualization builds upon binding and non-binding international human rights standards, such as the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD), the Convention on the Prevention and Punishment of the Crime of Genocide (Genocide Convention), the International Covenant on Civil and Political Rights (ICCPR), Article 20(2), the Decision on combating certain forms and expression of racism and xenophobia by means of criminal law (EUFD 2008/913/JHA), the case law of the ECtHR, and the European Commission against Racism and Intolerance (ECRI)'s General Policy Recommendation No. 7. As a result, Paragraph 11 can be claimed to represent international human rights standards.

This Chapter adopts a critical approach to international human rights by assuming an expansive interpretation of impermissible grounds of Paragraph 11 as the working definition for the following sections. To clarify, Paragraph 11 could have more clearly adopted an expansive conceptualization of the impermissible grounds[32] for hate speech, i.e. "racist, xenophobic, sexist and LGBTI-phobic".[33] The conceptualization of 'hate speech' by critical race scholars[34] highlights that hate speech is used to perpetuate systems of histor-

30 European Parliament (2023) Legislative Train Schedule, Proposals to extend the list of EU crimes to all forms of hate crime and hate speech, available at <https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-hate-crimes-and-hate-speech> accessed 7 Sep 2023.

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

32 Tarlach McGonagle 'Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas' (2011). This research employs 'impermissible grounds' as an expression that aims to emphasise the wrongful act and the perpetrator as opposed to focusing on the targeted groups. Additionally, this research avoids the expressions 'victims' or 'vulnerable groups' noting that people historically and systematically targeted by hate speech have criticised how such terms can be wrongfully interpreted as passive states of subjugation. 'Victims' may be used for legal coherence when referring to legal instruments such as the European Union Victims' Rights Directive 2012/29/EU.

33 Eva Nave, 'Hate speech, historical oppression and European human rights (2023 forthcoming) Buffalo Human Rights Law Review; Eva Nave and Lottie Lane, 'Countering online hate speech: How does human rights due diligence impact terms of service?' (2023) Computer Law & Security Review.

34 Critical race theory is the legal scholarship grounding the understanding and importance of a legal regime regulating 'hate speech' in reference to 'racist hate speech'. Mari J. 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'. Mari J Matsuda, 'Public Response to Racist Speech: Considering the Victim's Story' (1989) 87 Michigan Law Review 2320, 2335.

ical and systematic oppression. Similarly, black feminist scholars[35] emphasize the need to reflect on the intersectionality of systems of oppression. As a result, CM/Rec(2022)16 could have improved legal coherence with the critical legal scholarship had it clearly adopted an expansive interpretation of impermissible grounds, taking into account the intersectionality of historical and systematic systems of oppression. An expansive interpretation of impermissible grounds would unequivocally offer a stronger human rights regime for groups targeted by criminal hate speech on the basis of, e.g., gender identity, religion, and ableism.

Importantly, only the most severe cases of hate speech should be criminalized.[36] When assessing the severity of the hateful expression, the ECtHR typically reviews a set of variables which Rosenfeld describes as the 'contextual variables approach'.[37] These variables include: the content of the speech;[38] the political and social context at the time of the speech;[39] the intention of the speaker;[40] the speaker's status or role in society;[41] the reach and form of dissemination of the speech;[42] the imminence or likelihood that the speech leads, directly or indirectly, to harmful consequences;[43] the nature and size

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

36 CM/Rec(2022)16, Explanatory Memorandum, available at <https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=0900001680a6891e> accessed 7 Sep 2023, Para. 20.

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

38 *E.g.*, Goucha v. Portugal, App. No. 70434/12 (Mar. 22, 2016), https://hudoc.echr.coe.int/fre?i=001-161527; Feldek v. Slovakia, App. No. 29032/95 (October 12, 2001), https://hudoc.echr.coe.int/fre?i=001-59588; Ottan v. France, App. No. 41841/12 (July 19, 2018), https://hudoc.echr.coe.int/fre?i=001-182627.

39 *E.g.*, *id.*; Ceylan v. Turkey [GC], App. No. 23556/94 (July 8, 1999), https://hudoc.echr.coe.int/fre?i=002-6560; Beizaras & Levickas v Lithuania, App. No. 41288/15 (Jan. 14, 2020), http://hudoc.echr.coe.int/eng?i=001-200344.

40 *E.g.*, Jersild v. Denmark, App. No. 15890/89 (July 8, 1993).

41 *E.g.*, Incal v. Turkey, App. No. 22678/93 (June 9, 1998), https://hudoc.echr.coe.int/fre?i=001-58197, where the ECtHR noted that politicians enjoy a protected status, but concomitantly have heightened responsibilities in that they should avoid disseminating comments in their public speeches which are likely to foster intolerance; Feret v. Belgium, App. No. 15615/07 (July 16, 2009), https://hudoc.echr.coe.int/eng-press?i=003-2800730-3069797, where the ECtHR noted that politicians have the duty to refrain from using or advocating for racial discrimination.

42 *E.g.*, Gündüz v. Turkey, App. No. 35071/97 (Dec. 4, 2003), https://hudoc.echr.coe.int/fre?i=001-61522, where the ECtHR stated that live TV as not easy to reformulate or retract.

43 *E.g.*, Erbakan v. Turkey, App. No. 59405/00 (July 6, 2006), https://hudoc.echr.coe.int/?i=001-76234, where the ECtHR found there had been a violation of Article 10 because there was no proof of actual risk or imminent danger of the speech fomenting intolerance.

of the audience;[44] and the victims' perspective including its size, homogeneity, its historical oppression.[45] The ECtHR takes into account how these variables interplay and interfere with the individuals' right to private life[46] to determine the most severe cases of hate speech.[47]

This Chapter develops a framework for the online detection of incitement to violence in E2EE services targeting historically or systematically oppressed people. This conceptualization stems from CM/Rec(2022)16 and it includes incitement to commit genocide, crimes against humanity, war crimes, and threats (the latter only applicable to threats of physical offences or to violation of the right to life). The rationale behind this conceptualization relates to the analysis of harm deriving from E2EE communications. To clarify, noting that groups on E2EE services are typically composed of like-minded people, people targeted by hate speech in such conversations would not be directly harmed if not in the group. Contrarily, E2EE group chats compromise the human rights of people targeted by hate speech if inciting the users in the group to violence outside the E2EE environment.

4.2.2 Implications on end-to-end encrypted services

While online messaging and social media have had beneficial impacts,[48] there are, however, also new human rights concerns associated with these digital environments. One of the most challenging aspects is enforcing content moderation practices[49] that are compliant with human rights. Thus far, this balancing act has tilted towards digital environments with little to no filtering resulting in the rise of online hate speech. While online hate speech was initially documented in publicly accessible settings, in recent years, the dynamics of spread of online hate have shifted to more privacy-securing

44 *E.g.*, Vejdeland & Others v. Sweden, App. No. 1813/07 (May 9, 2012), https://hudoc.echr.coe.int/eng?i=001-109046; Vereinigung Bildender Künstler v. Austria, App. No. 68354/01 (April 25, 2007), https://hudoc.echr.coe.int/fre?i=001-79213.

45 *E.g.*, Leroy v. France, App. No. 36109/03, ¶ 27, 31, 43 (Oct. 2, 2008), https://hudoc.echr.coe.int/eng-press?i=003-2501837-2699727.

46 ECHR, Art. 8.

47 *E.g.*, Kiraly & Domotor v. Hungary, App. No. 10851/13 (April 17, 2017, https://hudoc.echr.coe.int/?i=001-170391, where the ECtHR found that authorities had failed to act against racial violence and breached the right to respect for private life under Article 8 ECHR.

48 *E.g.*, Gadi Wolfsfeld, Elad Segev, and Tamir Sheafer 'Social media and the Arab Spring: Politics comes first' (2013) Journal of Press/Politics 18(2): 115-137.

49 Content moderation refers to a set of policies, processes and digital technologies used by internet intermediaries to review user-generated content and to decide what content is to, in broad terms, remain or be removed from online environments.

environments.[50] In particular, hate mongers increasingly seek platforms offering the possibility of exchanging information through a specific type of encryption, *i.e.*, E2EE.[51]

E2EE services enable message communication between two (or more) users while ensuring that nobody else can access their content. This is achieved by encrypting and decrypting their messages with a cryptographic key that is only known to the two (or the group of) users. Typically, internet intermediaries providing the E2EE service do not have the cryptographic key, and do not access the content of the users' messages.[52]

E2EE services are provided by a wide range of internet intermediaries[53] such as: email services (e.g. ProtonMail, Tutanota, Thunderbird); video conferencing services[54] (e.g. Zoom, Skype, Google Meet, Microsoft Teams);[55] and – the most relevant for our article – messaging services (e.g. Signal, WhatsApp, Telegram, Viber,[56] Facebook Messenger,[57] Instagram[58]). These messaging services are provided by online platforms[59] which have adopted E2EE either by default or opt-in. Importantly, the engagement features in E2EE are expanding beyond one-on-one messaging. Online platforms such as WhatsApp and

50 *E.g.*, ABC News (2023) Donal Trump Supporters embrace Signal, Telegram and other 'free speech' apps, available at <https://www.abc.net.au/news/2021-01-20/donald-trump-social-media-apps-free-speech-privacy/13071206> accessed 7 Sep 2023, Foreign Policy (2021) Are Telegram and Signal Havens for Right-Wing Extremists? available at < https://foreignpolicy.com/2021/03/13/telegram-signal-apps-right-wing-extremism-islamic-state-terrorism-violence-europol-encrypted/#cookie_message_anchor> accessed 7 Sep 2023.

51 Tech against Terrorism (2021) Use of E2EE: State of Play, Misconceptions, and Mitigation Strategies, available at < https://www.techagainstterrorism.org/2021/09/07/terrorist-use-of-e2ee-state-of-play-misconceptions-and-mitigation-strategies/> accessed 7 Sep 2023, 42.

52 Fornetix (2022) End-to-End Social Media Encryption Strategies, available at <https://www.fornetix.com/articles/end-to-end-encryption-strategies-becoming-the-norm-for-social-media/> accessed 7 Sep 2023; Ben Lutkevich and Madelyn Bacon (2021) Definition end-to-end encryption (E2EE) available at <https://www.techtarget.com/searchsecurity/definition/end-to-end-encryption-E2EE> accessed 7 Sep 2023.

53 See supra (n 7).

54 Emily R (2022) Top 7 Most Secure Video Calling Apps, available at <https://getstream.io/blog/safest-video-calling-apps/> Accessed 7 Sep 2023.

55 Anina OT (2021) What Apps Use End-to-End Encryption to Improve Online Privacy, available at <https://www.makeuseof.com/apps-use-end-to-end-encryption/> accessed 7 Sep 2023. X Corp. direct messaging service will be E2EE, see Zoe Kleinman and Tom Gerken (2023) Twitter launches encrypted private messages, says Elon Musk, available at <https://www.bbc.com/news/technology-65533021> accessed 7 Sep 2023.

56 Anthony Spadafora (2023) The best encrypted messaging apps in 2023, available at <https://www.tomsguide.com/reference/best-encrypted-messaging-apps> accessed 7 Sep 2023.

57 Timothy Buck (2022) Update to End-to-End Encrypted Chats on Messenger, available at <https://about.fb.com/news/2022/01/updates-to-end-to-end-encrypted-chats-messenger/> accessed 7 Sep 2023.

58 Instagram Help Centre (2023) How do I start an end-to-end encrypted chat on Instagram, available at <https://help.instagram.com/1165835007222763/?helpref=related_articles> accessed 7 Sep 2023.

59 See supra (n 17).

Signal allow group communication up to 1000 users[60] and WhatsApp has built-in in-chat shopping options.[61]

E2EE services have both benefits and risks.[62] On the one hand, E2EE services preserve privacy and enable safer interaction between human rights activists.[63] On the other hand, the same privacy feature challenges accountability and attracts criminal activity. Moreover, given the large number of users allowed in groups on E2EE services, the likelihood and imminence of harm can be considered the highest when compared to other digital settings. For example, on open-ended encryption platforms, as content is publicly shared, it can be more frequently reported by other users and, ultimately, removed if illegal.

Ongoing strategies to counter illegal content, such as hate speech, on E2EE services are challenging human rights. Law enforcement bodies struggle to operationalise their mandate as hate mongers use E2EE services to hide their communications from public oversight. As a result, law enforcement may adopt practices that are not compliant with human rights such as, infiltration,[64] provocation,[65] or requests by of backdoors to access private communication.[66]

Similarly, internet intermediaries also struggle to provide their services without hosting online hate speech. Typically, platforms have relied on user reports of hate speech. However, considering that most groups using E2EE are composed of like-minded people, reporting is unlikely. Ongoing debates seek to conceptualize corporate HRDD responsibilities not to host illegal content, such as hate speech, in a way that does not disproportionately interfere

60 Signal Support, Group chats, available at <https://support.signal.org/hc/en-us/articles/360007319331-Group-chats#:~:text=Admin%20controls%20of%20who%20can%20send%20messages%20and%20start%20calls,Size%20limit%20of%201000> accessed 7 Sep 2023.

61 Ingrid Lunden (2020) facebook adds hosting, shopping features and pricing tiers to WhatsApp Business, available at <https://rb.gy/2sj7p> accessed 7 Sep 2023.

62 Maria Koomen, Carnegie Endowment for International Peace (2021) the Encryption Debate in the European Union: 2021 Update, available at <https://carnegieendowment.org/2021/03/31/encryption-debate-in-european-union-2021-update-pub-84217> accessed 7 Sep 2023.

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

64 Often disproportionally affecting marginalized communities. See *e.g.* Amnesty International (2017) Attacks on human rights activities reach crisis point globally, available at <https://www.amnesty.nl/actueel/attacks-on-human-rights-activists-reach-crisis-point-globally> accessed 7 Sep 2023; Ashely D. Farmer, Organization of American Historians, available at<https://www.oah.org/tah/history-for-black-lives/tracking-activists-the-fbis-surveillance-of-black-women-activists-then-and-now/> accessed 7 Sep 2023.

65 *E.g.*, Snow, D. Della Porta, D., Klandermans, B. and McAdam, D. (eds.) Encyclopedia of Social and Political Movements, Agents Provocateus as a Type of Faux Activist, available at <https://web.mit.edu/gtmarx/www/agentsprovocateursfaux.html> accessed 7 Sep 2023.

66 *E.g.*, following the terrorist attacks in San Bernardino in 2015 and Pensacola in 2019, the FBI requested backdoors to Apple's iPhone software, available at <https://en.wikipedia.org/wiki/End-to-end_encryption#Backdoors> accessed 7 Sep 2023.

with the rights to freedom of expression, assembly and association, privacy, and with data protection.

4.2.3 Key human rights safeguards in countering criminal hate speech in E2EE

This section analyses the main human rights safeguards in countering criminal hate speech on E2EE covering the operationalization of the legal requirements for restricting freedom of expression, assembly and association, data protection, and privacy rights (further analysed in Section 4.5.3).

4.2.3.1 Freedom of expression, assembly and association

The ECtHR has posited that freedom of expression applies "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."[67] The CM/Rec(2022)16 reinforced that interferences with the right to freedom of expression must be "construed narrowly".[68]

Article 10(2) prescribes that restrictions on the right to freedom of expression must 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.

Any restriction on the exercise of the right to freedom of expression must address a pressing social need and be proportional. This means that such restriction must be the least intrusive measure, whereby the protection of human rights outweighs the limits on freedom of expression.[69] Notwithstanding, the ECHR also prescribes that the exercise of the right to freedom of expression entails specific "duties and responsibilities" which when not respected may encompass legal restrictions.[70]

Article 11 ECHR sets out the right to freedom of assembly and association clarifying the "right to freedom of peaceful assembly and freedom of association with others".[71] Similarly to Article 10, also Article 11 foresees the

67 Handyside v. UK, App. No. 5493/72, ¶ 49 (Dec. 7, 1976), https://hudoc.echr.coe.int/eng?i=001-57499.
68 CM/Rec(2022)16, Explanatory Memorandum, Para. 48.
69 CM/Rec(2022)16, Explanatory Memorandum, Para. 48.
70 ECHR, Art. 10(2).
71 ECHR, Art. 11.

possibility of restrictions as long as they are: (i) prescribed by law; (ii) necessary in a democratic society; and (iii) in pursuit of legitimate interests such as national security or public safety, the prevention of disorder or crime, the protection of health or morals or for the protection of the rights and freedoms of others. Notably, the possibility for restricting the right to freedom of assembly and association also applies to governments and law enforcement bodies.[72]

4.2.3.2 Privacy and data protection

Countering criminal hate speech on E2EE services also requires compliance with the requirements emanating from both the right to respect for private and family life (broadly referred to as right to privacy) and the right to the protection of personal data (broadly referred to right to data protection).[73]

On the one hand, everyone has the right to privacy as per Article 8 of the ECHR, and Article 7 of the CFREU. These articles encapsulate the legal framework through which no one (including other individuals, private actors, or public bodies) has the right to know details about a person's life unless specifically provided by law. Further to this, the Directive on privacy and electronic communications (e-Privacy Directive)[74] supplements the protection of privacy in the context of the electronic communications sector. Article 5 prescribes the general confidentiality of electronic communications and the obligation for Member States to adopt national legislation that prohibits listening, tapping, storage or other kinds of interception or surveillance of communications and the related traffic data.[75] There are two exceptions to this obligation: (i) the users' consent and (ii) a legal authorisation according to Article 15.[76] The Court of Justice of the EU (CJEU) has ruled that the legal authorisation criterion must be interpreted in a restrictive manner, *i.e.* in accordance with "Member States law".[77] Applying the e-Privacy Directive framework to the research in this Chapter, E2EE services arguably have the HRDD responsibility to counter criminal hate speech as long as prescribed in the domestic legal frameworks in which they operate.

---

72 ECHR, Art. 11(2).

73 For further analysis see Gloria Gonzaìlez-Fuster, Rosamunde Van Brakel, and Paul De Hert (Eds.) Research handbook on privacy and data protection law: values, norms and global politics (2022) Edward Elgar Publishing; Paul De Hert and Serge Gutwirth 'Privacy, data protection and law enforcement. Opacity of the individual and transparency of power.' (2006) Privacy and the criminal law: 61-104.

74 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), L 201.

75 e-Privacy Directive, Art. 5.

76 e-Privacy Directive, Art. 5(1).

77 *E.g.,* Judgment of the Court (Grand Chamber) of 21 December 2016 Tele2 Sverige AB v Post- och telestyrelsen and Secretary of State for the Home Department v Tom Watson and Others, ECLI:EU:C:2016:970.

On the other hand, countering criminal hate speech on E2EE services involves the process of personal data as it comprises the processing of information related to an identifiable natural person as per Article 4 (1) of the General Data Protection Regulation (GDPR).[78] Although everyone has the right to protection of personal data,[79] the collection of personal data is possible as long as within legal limits. The right to data protection has different implications depending on the actor processing the personal data. If considering the process of personal data by the internet intermediaries, the GDPR applies. If considering the process of personal data by law enforcement, the Data Protection and Law Enforcement Directive applies.[80]

In this context, this Chapter focuses primarily on the HRDD of private actors and thus investigates more thoroughly the GDPR requirements.[81] Articles 5 and 6 of the GDPR state the data protection principles containing the legal bases for the processing of personal data. Article 5 lays down the principles for processing personal data which broadly include: lawfulness, fairness, and transparency; purpose limitation; data minimisation; accuracy; storage limitation; integrity and confidentiality; and, accountability.[82] Article 6 expands on the criteria needed to establish a lawful basis of processing which encompasses: consent given by the data subject for specific purposes; performance of a contract to which the data subject is party; compliance with a legal obligation; protection of vital interests of the data subject; performance of a task carried out in the public interest or in the exercise of official authority; legitimate interests pursued by the controller or a third party.[83]

Applying the GDPR framework to the research in this Chapter, E2EE services arguably have the HRDD responsibility to process personal data associated with countering criminal hate speech under four main legal bases. First, E2EE services have the legal obligation, in accordance with EU law or domestic law, to counter criminal hate speech.[84] Second, in the cases of imminence of harm, it may be necessary that E2EE process personal data to

78 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), L 119/1.

79 CFREU, Art. 8.

80 Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA

81 In this Chapter, the data protection requirements applicable to law enforcement are relevant in a later analysis in Section 5.2.

82 GDPR, Art. 5.

83 GDPR, Art. 6.

84 GDPR, Arts. 6(1)(c) and 6(3).

protect data subjects or another natural person.[85] For example, the case in which a mob is organizing on E2EE inciting physical harm or killing of someone or a group of people. Third, E2EE services may also have the legitimate interest that their services are safely provided.[86] Fourth, E2EE services may have the data subject's consent as long as users are adequately informed about the specific purpose and circumstance for the data processing.[87]

Section 4.2. clarified that incitement to violence is one of the most serious forms of hate speech which should be criminalised and prohibited on online environments, such as in E2EE. Additionally, this section explained that measures to counter criminal hate speech in E2EE must comply with minimum human rights safeguards.

4.3 Corporate human rights due diligence (HRDD) to counter hate speech in E2EE

Though current legislation creates corporate HRDD responsibilities to counter online hate speech, due to insufficient interdisciplinary debate, the HRDD regime has not been properly expanded to E2EE services. The HRDD framework covers preventive, promotional and remedial responsibilities. The applicable HRDD framework depends on the type and size of the internet intermediary. The extent to which HRDD should be implemented depends on technological advancements (Section 4.4).

4.3.1 Internet intermediaries' responsibility to protect human rights

The general corporate responsibility to protect human rights is articulated in legal standards both at the international and at the European level. At the international level, the United Nations Guiding Principles on Business and Human Rights (UNGPs) is the most influential instrument.[88] Though not binding, the UNGPs were unanimously endorsed by the UN Human Rights Council in 2011 and are the universal frame of reference for the businesses' responsibility to prevent and mitigate human rights abuses.

Businesses should have in place policies and processes to respect human rights including: '(a) a policy commitment to meet their responsibility to respect human rights; (b) a HRDD process to identify, prevent, mitigate and account for how they address their impacts on human rights; (c) processes to enable the remediation of any adverse human rights impacts they cause or to which

85 GDPT, Art. 6(1)(d).
86 GDPR, Art. 6(1)(f).
87 GDPR, Art. 6(a).
88 UNGPs (n 18)

they contribute.'[89] Notably, the policy commitment should be publicly available and communicated to all stakeholders associated with its operations and potentially affected by human rights abuses.[90] The HRDD process places an emphasis on preventive responsibilities, as businesses should "(a) *avoid* causing or contributing to adverse human rights impacts through their own activities (…), and (b) seek to *prevent or mitigate* adverse human rights impacts that are directly linked to their operations, products, or services by their business relationships, even if they have not contributed to those impacts."[91]

At the EU level, two instruments would expand the HRDD framework. First, at a cross-sector level, the CSDDD. The remit of its application is threefold: (1) EU companies with 500+ employees and a turnover of over _150 million worldwide; (2) non-EU companies with an equivalent turnover threshold generated in the EU;[92] and (3) companies falling outside this remit of application but operating in "high-impact sectors" are also required to follow the HRDD framework in the CSDDD.[93]

Companies within the scope of the CSDDD, including those providing E2EE services, must adopt a HRDD framework to identify, prevent, mitigate, and account for their adverse impacts on human rights[94] throughout their operations and value chains.[95] The human rights conceptualisation in the CSDDD includes instruments covering criminal hate speech in relation to incitement to violence. Relevant instruments include the Genocide Convention, ICERD, and ICCPR. Relevant provisions include the right to life and security,[96] violation of the prohibition of torture, cruel, inhuman or degrading treatment.[97] Arguably, such HRDD framework applies to E2EE services provided by very large platforms such as Facebook (Messenger),[98] WhatsApp.[99] Regrettably, the turnover threshold in the CSDDD leaves many impactful online services

89 UNGPs, Principle 15.
90 UNGPs, Principle 16.
91 UNGPs, Principle 13.
92 CSDDD, General Approach, Art. 1.
93 CSDDD, General Approach, Recitals 21-23, 15. The current draft does not include social media companies as high-impact sector companies.
94 CSDDD, Explanatory Memorandum, 3.
95 CSDDD, Explanatory Memorandum, 3.
96 Universal Declaration of Human Rights (adopted 10 December 1948) 217 A(III) (UNGA) (UDHR), Art. 3; International Covenant on Civil and Political Rights (adopted 16 December. 1966, entered into force 23 March 1976) 999 UNTS 171 (ICCPR), Art. 6.
97 UDHR, Art. 5, ICCPR, Art. 7.
98 Mansoor Iqbal (2023) Facebook Revenue and Usage Statistics, available at <https://www.businessofapps.com/data/facebook-statistics/> accessed 7 Sep 2023, reports a turnover of 116 billion USD.
99 Mansoor Iqbal (2023) WhatsApp Revenue and Usage Statistics, available at <https://www.businessofapps.com/data/whatsapp-statistics/> accessed 7 Sep 2023, reported a turnover of 906 million USD.

outside the mandatory preventive HRDD regime, including E2EE services involved in the rise of hate mongers such as Telegram.[100]

The second European instrument expanding the HRDD framework is the AI Act. The AI Act[101] introduces sector-specific HRDD responsibilities for companies using AI systems based on three risk levels: unacceptable risk AI; high-risk AI; low or minimal risk AI.[102] The EP Compromise Amendments suggests that social media companies[103] be considered high-risk, however only with respect to their recommender systems.[104]

As such, in its current form, the AI Act HRDD framework does not seem to apply to E2EE services. Nevertheless, the monetisation of E2EE services with shopping features, such as WhatsApp, raises the question of whether the online platforms will conduct any type of content regulation equivalent to link-recommendation, in which case the AI Act HRDD regime would apply.

4.3.2 Corporate HRDD to counter criminal hate speech online

This section covers the main corporate HRDD regimes in Europe applicable to online platforms in countering online hate speech.[105] The DSA sets the goals and means to achieve the harmonisation of intermediary liability and HRDD rules to protect the rights in the CFREU.[106] This Chapter focuses on the elements of HRDD within the DSA.

The DSA HRDD responsibilities are tailored for different internet intermediaries, depending on their role, size, and impact.[107] The HRDD regime

100 Mansoor Iqbal (2023) Telegram Revenue and Usage Statistics, available at <https://www.businessofapps.com/data/telegram-statistics/> accessed 7 Sep 2023.

101 As the AI Act is a Regulation, the goals and the means to achieve said goals are binding on all EU MS.

102 AI Act, 3.

103 Meaning equivalent to online platforms.

104 European Parliament, Draft Compromise Amendments on the Draft Report, AIA, KMB/DA/AS, version: 1.1 available at < https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf > accessed 7 Sep 2023, Title III, Chapter 1, Annexes II and III and recitals 27 to 41a, 40b.

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

106 DSA, Art. 1(1).

107 DSA, Recital 41.

applicable to internet intermediaries can be broadly subdivided in the following pyramidal structure: on the base, HRDD responsibilities of all internet intermediaries; in the middle, HRDD responsibilities of hosting services, including online platforms; at the top, HRDD responsibilities of very large online platforms (VLOPs) and very large online search engines (VLOSEs). The DSA complements the AVMSD which prescribes HRDD responsibilities for video-sharing platforms.[108]

Internet intermediaries have the general preventive HRDD responsibilities to, upon knowledge, expeditiously remove illegal content on its service,[109] and to design terms of service (ToS) compliant with fundamental rights, namely complying with the prohibition of hate speech.[110] Though hate speech is considered illegal content in EU law,[111] the legal conceptualisations of impermissible grounds for hate speech vary depending on the instrument.[112] This Chapter adopts an extensive conceptualisation of hate speech grounded in an analysis of historical and intersectional systems of oppression.

The DSA does not allow for a general monitoring obligation to detect illegal content,[113] but it does mention the possibility of having specific monitoring obligations imposed on internet intermediaries "by national authorities in accordance with national legislation, in compliance with Union law(…)".[114] Additionally, hosting services, including online platforms, must also notify law enforcement if they suspect that a criminal offence involving a threat to the life or safety of a person has taken place, is taking place or is likely to take place.[115] Within the scope of online platforms, the DSA creates heightened HRDD for platforms with higher risks due to their larger reach and impact, *i.e.* companies with 45 million or more average monthly active recipients of their service in the Union, referred to as VLOPs and VLOSEs.[116]

VLOPs and VLOSEs must "*diligently* identify, analyse and assess systemic risks",[117] which include *inter alia* the dissemination of illegal content and any actual or foreseeable negative effects for the exercise of fundamental rights,

108 The DSA is complementary to the AVMSD.
109 DSA, Art. 6(1)(b).
110 DSA, Art. 14(4), see Naomi Appelman, João Pedro Quintais and Ronan Fahy, 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation' (2022) German Law Journal.
111 European Commission, Recommendation 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, L 63/50.
112 Eva Nave (n 33), Eva Nave and Lottie Lane (n 33), Natalie Alkiviadou, The Legal Regulation of Hate Speech: The International and European Frameworks, 55 Politicka Misao 203, 223 (2018).
113 DSA, Recital 30.
114 DSA, Recital 30.
115 DSA, Art. 18(1).
116 DSA, Recitals 57 and 76.
117 DSA, Art. 34. Emphasis added.

such as human dignity,[118] respect for private and family life,[119] protection of personal data,[120] freedom of expression and information,[121] and non-discrimination.[122] Mitigation measures to address these systemic risks include adapting ToS, disabling access to the content in particular in respect to illegal hate speech or cyber violence, and cooperating with other providers through codes of conduct or crisis protocols.[123]

Applying the DSA HRDD framework to E2EE services, the latter fall within the category of internet intermediaries either as a i) 'mere conduit' transmitting in a communication network information provided by the user, or providing access to a communication network or ii) a 'hosting' service storing information provided by and at the request of the user. Most E2EE services would qualify as internet intermediaries under i), yet, in certain cases such as WhatsApp Businesses, it would also qualify as internet intermediaries under ii). Furthermore, given that Recital 20 extends the intermediary liability exemption regime in the DSA to internet intermediaries providing encrypted transmissions, one can logically assume that the HRDD framework for internet intermediaries also applies to internet intermediaries using E2EE services. Some E2EE services may also fall under the definition of online platform if catering to a public groups or open channels,[124] as could arguably be the case of E2EE chats allowing for public groups and open channels.[125] Additionally, online platforms and VLOPs may also provide E2EE services in their messaging applications, such as Facebook Messenger.[126]

The 2018-revised AVMSD also imposes HRDD responsibilities for audiovisual media services as TV broadcasters, video-on-demand services, and video-sharing platforms.[127] Video-sharing platforms are defined as platforms providing programmes or user-generated videos to the general public with the purpose of entertaining or educating.[128] The video-sharing platform must algorithmically organize the videos by displaying, tagging, and sequenc-

118 CFREU, Art. 1 .
119 CFREU, Art. 7.
120 CFREU, Art. 8.
121 CFREU, Art. 11.
122 CFREU, Art. 21.
123 DSA, Art. 53(1).
124 DSA, Recital 14.
125 Examples available at <https://www.whtsgrouplinks.com/> accessed 6 Feb 2024.
126 Facebook Help Center, What end-to-end encryption on Messenger means and how it works, available at <https://www.facebook.com/help/messenger-app/786613221989782?cms_id=786613221989782> accessed 7 Sep 2023.
127 European Commission, Guidelines on practical application of the essential functionality criterion of the definition of a 'video-sharing platforms service' under the Audiovisual Media Services Directive (2020/C 223/02), C 223/3 available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.C_.2020.223.01.0003.01.ENG&toc=OJ:C:2020:223:TOC> accessed 7 Sep 2023, I. Introduction citing Article 1(1)(aa) of the AVMSD.
128 AVMSD, Art. 1(aa).

ing.[129] The AVMSD prescribes heightened HRDD responsibilities for video-sharing platforms, requiring these to explicitly refer in their terms of service the prohibition of hate speech. Notably, the AVMSD follows the expansive interpretation of impermissible grounds in Article 21 CFREU.[130]

Applying the HRDD framework in the AVMSD to E2EE services, there are two aspects to consider. First, should the "general public" element be interpreted as to refer to a large audience, given the current features in some E2EE services allowing public groups and open channels, E2EE services with these features should fall under the definition of general public in the AVSMD. Second, though typically there is no editorial responsibility in E2EE communication services be it in messaging, videos, or e-mail, the Graphics Interchange Format (GIF)[131] features in such applications do include some type of content curation by the platforms. The growing use of GIFs by hate mongers[132] requires legal framing, and one possible way could be through the AVMSD.

The Code of Conduct to counter illegal hate speech online (CoC) is a co-regulatory instrument signed in 2016 as an agreement between the European Commission and some of the largest internet intermediaries. Originally, Meta Platforms, Inc. (previously Facebook, Inc.), Microsoft, X Corp. (previously Twitter, Inc.) and YouTube; over time, Instagram, Snapchat, Dailymotion, Jeuxvideo, TikTik, LinkedIn, Rakuten, Viber and Twitch also became part of the CoC.[133] The CoC emphasises preventive HRDD responsibilities to counter incitement to violence and hateful conduct that include: clarity and transparency in the drafting of the ToS; improvement of mechanisms for notices, flagging, and review of said content; education and awareness-raising initiatives with users and staff; and collaboration with civil society acting as trusted flaggers.

The CoC applies to the E2EE services provided by the signatory companies such as Facebook Messenger, Snapchat, Viber, and the recently-launched X Corp. encrypted messaging feature.[134] However, in the monitoring reports of the CoC there is no mention of how companies should implement HRDD in their E2EE services.

At the CoE level, the CM/Rec(2022)16 is a key standard-setting policy instrument clarifying the that internet intermediaries must comply with HRDD

129 AVMSD, Art. 1(aa).

130 Eva Nave and Lottie Lane (n 33).

131 A GIF is a bitmap image format that also supports animations.

132 Khosravi Ooryad, S. (2023). Alt-right and authoritarian memetic alliances: global mediations of hate within the rising Farsi manosphere on Iranian social media. *Media, Culture and Society*. https://doi.org/10.1177/01634437221147633, 498.

133 CoC (n 23).

134 Siladitya Ray (2023) Encrypted Messaging, 2-Hour Videos: Here Are the Moves Twitter Has Made in Its Bid To Become an 'Everything' App, available at <https://www.forbes.com/sites/siladityaray/2023/05/26/encrypted-messaging-2-hour-videos-here-are-the-moves-twitter-has-made-in-its-bid-to-become-an-everything-app/> accessed 7 Sep 2023.

responsibilities, including with legislation on hate speech.[135] It specifies that internet intermediaries must *inter alia*: explicitly state in their terms of service how they align with human rights;[136] remove the most severe cases of hate speech i.e. criminal hate speech;[137] and, report to public authorities criminal hate speech.[138] The HRDD responsibility to report criminal law to public authorities is aimed at facilitating investigations and remediation processes. To assess the severity of the hate speech and to design appropriate and proportionate countering measures, CM/Rec(2022)16 clarifies that all stakeholders, including States and its law enforcement actors as well as internet intermediaries alike, should assess the contextual variable (Section 4.2.1).[139]

The standards in the CM/Rec(2022)16 apply to internet intermediaries "regardless of their size, sector, operational context, ownership structure, or nature".[140] Nevertheless, this Recommendation explains that the means to address online hate speech "should be calibrated according to the severity of the human rights impact".[141] The CM/Rec(2022)16 aligns with the approach adopted by the DSA and prescribes stronger HRDD responsibilities for internet intermediaries comprising higher risk of contributing to human rights abused. Hence, given the heightened human rights risk of sharing criminal hate speech in E2EE application, internet intermediaries providing E2EE applications should consider adopting "greater precautions".[142]

4.3.3 Corporate HRDD to counter illegal content in E2EE services

There is currently no specific legislation regulating the HRDD responsibilities to counter online hate speech of internet intermediaries providing E2EE services. This section reviews two regulatory instruments impacting the HRDD responsibilities of E2EE services in the context of two different types of illegal content *i.e.*, terrorism (Section 4.3.3.1) and child sexual abuse material (Section 4.3.3.2).

4.3.3.1 EU Regulation on Terrorist Content Online

The EU Regulation on Terrorist Content Online (TCOR), in force since 2021, obliges hosting service providers to take proactive measures to prevent the dissemination of terrorist content and to respond within one hour to orders

135 CM/Rec(2022)16, Para. 18.
136 CM/Rec(2022)16, Para. 31.
137 CM/Rec(2022)16, Para. 31.
138 CM/Rec(2022)16, Para. 2.2.
139 CM/Rec(2022)16, Explanatory Memorandum, Para. 34.
140 CM/Rec(2022)16, Explanatory Memorandum, Para. 124.
141 CM/Rec(2022)16, Explanatory Memorandum, Para. 124.
142 CM/Rec(2022)16, Explanatory Memorandum, Para. 128.

issued by law enforcement bodies to remove such content.[143] 'Hosting service providers' covers providers storing information and making it available at the request of the user to other users,[144] thus including social media, video, image, and audio-sharing services. The TCOR applies to all platforms, regardless of size, as long as it has a significant number of users in one or more EU MS,[145] and it imposes fines on non-compliant companies.[146] Notably, the TCOR specifically incentivises hosting service providers to proactively remove content containing imminent life threats.[147]

The TCOR has been criticised for not setting enough human rights safeguards. Firstly, it not only adopts a vague conceptualisation of 'terrorist content', but it also allows providers to decide which automated content regulation algorithms to use.[148] Secondly, removal orders can be issued by entities that will not decide in an impartial way.[149] Thirdly, the one hour timeframe for all providers is likely to disproportionally hinder smaller businesses.

4.3.3.2 EU Proposal Regulation on Child Sexual Abuse Material[150]

Currently, the EU allows for internet intermediaries providing messaging and e-mail services to voluntarily use technologies to process personal data and

143 'Terrorist content' is defined as acts that 'seriously intimidate a population, unduly compelling a government or an international organisation to perform or abstain from performing any act, seriously destabilising or destroying the fundamental political, constitutional, economic or social structures of a country or an international organisation, TCOR, Art. 3.

144 TCOR, Art. 2(1).

145 TCOR, Art. 2.

146 TCOR, Art. 18.

147 TCOR, Art. 3.

148 European Digital Rights (EDRi) (2022) A safe internet for all, Upholding private and secure communication, available at <https://edri.org/wp-content/uploads/2022/10/EDRi-Position-Paper-CSAR.pdf> accessed 7 Sep 2023, 24 and 25.

149 EDRi (n 148), 59.

150 For illustrative purposes of the potential legal challenges, this thesis presents the analysis as originally published in the academic journal, which is based on the 2022 proposal by the European Commission for a Regulation to prevent and combat CSAM, *i.e.* European Commission (2022) COM (2022) 209: Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, available at <https://eur-lex.europa.eu/legal-content/EN/HIS/?uri=CELEX:52022PC0209> accessed 7 Sep 2023. It should be noted that the European Commission published, in 2024, a new Proposal for a Directive to prevent and combat CSAM, *i.e.* European Commission (2024) COM (2024) 60 final 2024/0035 (COD): Proposal for a Directive of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse and replacing Council Framework Decision 2004/68/JHA (recast) available at <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2024%3A60%3AFIN> accessed 20 November 2024.

other data to the extent necessary to detect, report, and remove child sexual abuse material (CSAM).[151]

The EU Proposal for a Regulation on Child Sexual Abuse Material (CSAR), proposed in 2022 by the European Commission, aims to harmonise objectives and implementation strategies on HRDD and liability regimes of internet intermediaries to identify, protect, and support victims of CSAM.[152]

The CSAR establishes risk assessments and mitigation frameworks complementary to those in the DSA.[153] The CSAR foresees the establishment of a 'Coordinating Authority' which, aside from overseeing the risk assessment and the subsequent mitigation measures put in place by the internet intermediaries, can also request a judicial or administrative authority to issue a detection order. Such a detection order results in specific mandatory obligations for the internet intermediaries to utilise digital technologies to detect the specific CSAM at the risk of receiving a fine up to 6% of its annual income or global turnover.[154]

The CSAR has been critiqued for negatively impacting data protection rights in two ways.[155] First, since it applies not only to cases of "known CSAM" but also to "child grooming" and other "new" material,[156] it is unclear what technological method could detect such content in a privacy protecting manner. To clarify, the CSAR seems to require the training an algorithm to detect new CSAM. In this regard, EDRi alerted to the low accuracy level of such an algorithm and hence to the lack of human rights

151 The e-Privacy Directive prevented internet intermediaries, including number-independent inter-personal communication services (NIICS) such as messaging services and email, from voluntarily using specific technologies to detect online CSA without authorization by national or EU legislation. On 2 August 2021, given the lack of EU legislation on CSAM, the EC adopted a temporary derogation to the e-Privacy Directive to allow for voluntary detection practices to continue. This regime is applicable until 3 August 2024 or until the CSAR is adopted.

152 CSAR (n 24), Explanatory Memorandim, 1.

153 CSAR (n 24), 2.

154 CSAR, Art. 35(2).

155 Ashel Smith (Bits of Freedom, 2022) European Commission wants to eliminate online confidentiality, available at <https://www.bitsoffreedom.nl/2022/05/11/european-commission-wants-to-eliminate-online-confidentiality/> accessed 7 Sep 2023; Jon Brodkin (2022) "War upon end-to-end encryption": EU wants Big tech to scan private messages, available at <https://arstechnica.com/tech-policy/2022/05/war-upon-end-to-end-encryption-eu-wants-big-tech-to-scan-private-messages/> accessed 7 Sep 2023.

156 A similar debate happened in the USA when Apple introduced two strategies to counter CSAM: messages notifying parents when children under 18 view CSAM and scans on iCloud Photos for CSAM to be then reported to Apple moderators. Both strategies were strongly criticised: Adi Robertson (the Verge, 2021) Apple's controversial new child protection features, explained, available at <https://www.theverge.com/2021/8/10/22613225/apple-csam-scanning-messages-child-safety-features-privacy-controversy-explained> accessed 7 Sep 2023.

safeguards.[157] For example, such algorithm would most likely detect consensual sexting between minors or adults looking like minors which would result in major privacy violation.[158] Arguably, though the European Commission prescribes that internet intermediaries should use the least privacy-intrusive method, the choice of method is left to the company's decision which does not guarantee human rights safeguards.

Second, the CSAR does not extend the voluntary detection currently in place. Instead, it instructs internet intermediaries to wait to receive a CSAM detection or removal order from judicial or administrative authorities. EDRi argued that 'such orders should only be issued by a court' to avoid having orders issued by for example judicial authorities such as prosecutors which in many member states are not independent authorities.[159]

In summary, ongoing proposals regulating HRDDR in E2EE services to counter illegal content fail to understand the digital technological possibilities and implications, and lack legal clarity and human rights safeguards.[160]

4.4 Digital technologies: content moderation in E2EE

This section expands on the digital technologies and encryption features used for content moderation[161] in E2EE services. Examples of content moderation methods in E2EE include: user reporting; message franking; message traceability; metadata analysis; perceptual hashing; private membership compu-

157 James Vincent (The Verge, 2022), New EU riles would require chat apps to scan private message for child abuse, available at <https://www.theverge.com/2022/5/11/23066683/eu-child-abuse-grooming-scanning-messaging-apps-break-encryption-fears?scrolla=5eb6d68b7fedc32c19ef33b4> accessed 7 Sep 2023.

158 Sabine Witting and Mark Leiser 'Outcome Reports of 1st expert Workshop on Eu proposed Regulation on Preventing and Combatting Child Sexual Abuse (2023) Council of Europe, available at <https://rm.coe.int/outcome-report-of-the-expert-workshop-on-eu-proposed-regulation-on-pre/1680aa00e4> accessed 17 October 2022; Sabine Witting and Mark Leiser 'Outcome Report of 2nd Expert Workshop on EU proposed Regulation on Preventing and Combatting Child Sexual Abuse (2023) Leiden University, available at <https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/final-eu-workshop-report-csa-proposal-2nd-workshop-05042023.pdf>accessed 17 October 2023.

159 EDRi (n 148).

160 Another example of an approach to counter CSAM lacking legal clarity is the Internet Watch Foundation (WF). The IWF supplies partner internet intermediaries with URLs that supposedly contain CSAM and should therefore be blocked. The IWF has been criticized for being ineffective and for lacking legitimate mandate. See CJ Davies (The Wired, 2009) The hidden censors of the internet, available at <https://www.wired.co.uk/article/the-hidden-censors-of-the-internet> accessed 5 Feb 2024, and Emily B. Laidlaw (2012) The responsibilities of free speech regulators: an analysis of the Internet Watch Foundation, *International Journal of Law and Information Technology*, https://doi.org/10.1093/ijlit/eas018

161 Though referred to as "content moderation techniques", this research acknowledges these techniques could also be referred to as content detection techniques.

tation; predictive models; multiparty computation.[162] This section focuses on metadata, hashing, combined with homomorphic encryption, as these ground the corporate HRDD responsibility standard proposed in this Chapter (Section 4.5.) to counter incitement to violence in E2EE.

4.4.1 Metadata

Metadata can be referred to as "data about data" and it can include file size, file type, date/time of creation or access, location, last modified field, sender/receiver, etc., without revealing the content of the message.[163] These types of metadata can be used to train machine learning models essentially in two ways. First, metadata such as data on the profile details can be used to predict the probability of having a user sharing CSAM on E2EE services.[164] Second, metadata such as data on the account creation activity, average shared messages or reports from other users, can be used to train machine learning models to predict a user's activity. These predictions can, supposedly, indicate the probability of a given user sharing illegal content like CSAM.[165] WhatsApp has acknowledged using metadata to predict the posting of CSAM.[166]

There are however significant human rights concerns regarding the use of metadata in E2EE services. On the one hand, the use of metadata can lead to the removal of legal content. For example, when used to classify spam or illegal content solely by monitoring the size or volume of the messages.[167]

On the other hand, there are also privacy concerns with metadata such as the identification of the sender and receiver.[168] A human rights safeguard

---

162 Center for Democracy &Technology (2021) Outside looking In – Approaches to Content Moderation in End-to-End Encrypted Systems, available at <https://cdt.org/wp-content/uploads/2021/08/CDT-Outside-Looking-In-Approaches-to-Content-Moderation-in-End-to-End-Encrypted-Systems-updated-20220113.pdf> accessed 7 Sep 2023; Chaintanya Rahalkar and Anushka Virgaonkar (2022) SoK: Content Moderation Schemes in End-to-End Encryption Systems, available at <https://click.endnote.com/viewer?doi=10.48550%2Farxiv.2208.11147&token=WzM2Njc3MjgsIjEwLjQ4NTUwL2FyeGl2LjIyMDguMTExNDciXQ.pz4XpiQvugO9XkR1TIhcQhsLW5I> accessed 7 Sep 2023; Sarah Scheffler and Jonathan mayer (2023) SoK: Content Moderation for End-to-end Encryption, available at <https://arxiv.org/pdf/2303.03979.pdf> accessed 7 Feb 2024.

163 Center for Democracy & Technology (n 162).

164 Center for Democracy & Technology (n 162), 21.

165 Center for Democracy & Technology (n 162), 21.

166 WhatsApp Help Center – How WhatsApp Helps Fight Child Exploitation. available at <https://faq.whatsapp.com/general/how-whatsapp-helps-fight-child-exploitation/?lang=en> accessed 7 Sep 2023.

167 Center for Democracy & Technology (n 162), 21; ; Chaintanya Rahalkar and Anushka Virgaonkar (n 162).

168 Greschbach, B., Kreitz, G., & Buchegger, S. (2012). The devil is in the metadata – New privacy challenges in Decentralised Online Social Networks. 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, 333–339, available

in this regard would be to regulate the use of metadata analysis to data that would not identify or would not so easily identify the user.

In the case of detecting incitement to violence in E2EE services allowing for the creation of public groups and open channels, metadata could be human rights compliant if regulated and used restrictively. This Chapter claims, first, that it is important to regulate which type of metadata service providers can access depending on which types of content they are trying to detect. Second, in the case of incitement to violence, the imminence of harm would increase with a rising number of users in a given group. Thus, it would arguably be proportionate to use metadata to identify large groups and to apply specific legal thresholds for content detection in such communities. The users would need to be effectively informed in the terms of service about these content detection thresholds applied to groups for the prevention incitement to violence.

4.4.2 Hashing

Hashing is a technique used to create a digital fingerprint (or "hash") for a given content to facilitate the matching of identical or similar content. There are two types of hashing techniques: cryptographic hashing and perceptual hashing.[169] Cryptographic hashing creates a random hash using a cryptographic function and it is usually used to identify known content without alterations. Perceptual hashing enables the identification of content up to a limited degree of differences. This technique is relevant to identify content with minor changes.

The detection of hashes at scale has been operationalised through the creation of databases where service providers share hashes of previously identified content. For example, CSAM, and terrorist content databases are already widely in use across the messaging services platforms.[170] Additionally, platforms may create databases of hashes for detecting specific content that they do not allow based on their ToS as is the case of Facebook's hashing database for intimate images non-consensually shared.[171] Importantly, detect-

at <https://doi.org/10.1109/PerComW.2012.6197506> accessed 7 Sep 2023, cited in Center for Democracy & Technology (n 162), 21.

169 For an overview see Center for Democracy & Technology (n 162), 22.

170 Center for Democracy & Technology (n 162), 22.

171 Meta (2019) Detecting Non-Consensual Intimate Images and Supporting Victims, available at <https://about.fb.com/news/2019/03/detecting-non-consensual-intimate-images/> accessed 7 Sep 2023.

ing content using perceptual hashing techniques is the most effective when content has been shared repetitively.[172]

In E2EE services, the scanning for the hashed content can happen at the server or client level, each encompassing different human rights risks. Scanning from the server's side can result in revealing information about the user to the server and thus may compromise privacy. Scanning from the client's side may be privacy compliant as long as the outcome of the scanning is not shared with the server.[173] It does however encompass a different problem which is that by revealing to the client the hash dataset, the client may then more easily circumvent it.[174] Additionally, client scanning may also raise more practical considerations as it would require that the user's device has a specific processing power, storage, internet connectivity, and battery capacity. This can disproportionately affect low-income individuals with low-end smartphones, or lead to individuals using low-end smartphones with the purpose of not performing the data processing.[175]

In the case of detecting incitement to violence in E2EE services, perceptual hashing from the client's side would potentially be human rights compliant. First, the users would have been informed in the terms of service about the use of specific content moderation techniques for the detection of incitement to violence in large group chats. In this context, the hash set containing the list of content classified as incitement to violence would be shared in the terms of service with the users. There is a risk of having users adjusting their behaviour and bypassing the hashing model by simply using a linguistic code avoiding the words categorized as incitements to violence. However, ultimately, any legal system must be clear and foreseeable.[176]

---

172 Interestingly, this was found to not be a very effective content detection technique in the case of CSAM as images reported are often new compared to the database of hashed content. See Bursztein, E., Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K., & Bright, T. (2019). Rethinking the Detection of Child Sexual Abuse Imagery on the Internet. The World Wide Web Conference, 2601–2607 available at <https://doi.org/10.1145/3308558.3313482> accessed 7 Sep 2023, cited in Center for Democracy & Technology (n 162).

173 Center for Democracy & Technology (n 162), 22. See also Sarah Scheffler, Anunay Kulshrestha, and Jonathan Mayer (2023) Public Verification for Private Hash Matching, available at <https://eprint.iacr.org/2023/029.pdf> accessed 7 Feb 2024.

174 Additionally, when the client does not know this dataset, they could easily forge the hash, thus avoiding detection.

175 James, J. (2020). The smart feature phone revolution in developing countries: Bringing the internet to the bottom of the pyramid. The Information Society, 36(4), 226–235 available at <https://doi.org/10.1080/01972243.2020.17614 97> accessed 7 Sep 2023, cited in Center for Democracy & Technology (n 162), 22.

176 There is however also the risk of abuse of a hashing solution by for example a governmental body which, instead of using a list of hashes that reflect incitement to violence, could use a list of hashes persecuting content displaying opposing political views. This Chapter emphasizes that this potential abuse must be prohibited and such a prohibition carefully enforced by a monitoring body.

Second, incitement to violence derives from a concrete legal framework which could be transformed into a hash set. Contrarily, CSAM cannot be summarized in a hash set, as CSAM content is different for each targeted child. In this context, a potentially privacy-preserving solution for CSAM detection would require the victim's self-identification and consent for hashing the abusive content for detection and further removal.

4.4.3 Homomorphic encryption

Homomorphic encryption is a form of encryption that enables an analysis of encrypted data without having to decrypt it first. The significant difference between this technique and traditional encryption methods is that, whilst the latter services had to decrypt the data to investigate it, with homomorphic encryption data can remain confidential while being processed and analysed.[177]

Depending on the type of mathematical computations (addition, multiplication or both) and whether these computations can be performed a limited or unlimited number of times, homomorphic encryption takes different forms: partially homomorphic encryption; somewhat homomorphic encryption; and Fully Homomorphic Encryption (FHE).[178] FHE is of special interest to our article as it enables all mathematical computations any number of times.

Typically, homomorphic encryption is useful for providers to perform operations on data that is stored or being transmitted as it avoids decryption during such operations and ensures data security. Common applications of FHE include securing data stored in the cloud, enabling data analytics in regulated industries (such as information technology), and improving election security and transparency. The main limitations to FHE are the difficulty to support multiple users and running complex algorithms. Nevertheless, some of the very large internet intermediaries like Google and Microsoft have started to implement and make homomorphic encryption available.[179]

In the case of detecting incitement to violence in E2EE services, homomorphic encryption can be of use as it enables the operationalisation of machine learning models in a privacy preserving manner. Thus, it can be combined with machine learning (in case of new unclassified content) or perceptual hashing (in case of known classified images) models for the identification of data archived, stored, or in transmission in the context of groups on messaging E2EE services. This technology appears to present the needed human rights

177 Anastasios Arampatzis (2023) Homomorphic Encryption: What Is It and How Is It Used, available at <https://venafi.com/blog/homomorphic-encryption-what-it-and-how-it-used/> accessed 7 Sep 2023.

178 Anastasios Arampatzis (n 177).

179 Anastasios Arampatzis (n 177).

safeguards for detection of incitement to violence in E2EE services. Nevertheless, given that this is a new digital technology, further research on the implementation at large scale is required.

4.5 Standard proposal: expanding HRDD to counter incitement to violence in E2EE services

This section proposes a legal standard expanding preventive and mitigatory HRDD responsibilities to counter incitement to violence in E2EE services by elaborating on the substantive regulation framework (Section 4.5.1), the procedural regulation (Section 4.5.2), the legal basis (Section 4.5.3), and the compliance with human rights safeguards (Section 4.5.4). The proposed HRDD standard can be summarised as a corporate HRDD responsibility to disrupt large groups inciting violence on E2EE.

4.5.1 Substantive regulation: Incitement to violence

According to European human rights standards, criminal hate speech covers a spectrum of acts ranging from incitement to genocide, incitement to violence, incitement to discrimination, threats, or insults (Section 4.2.1). This Chapter proposes a HRDD standard that applies to the acts of incitement to violence.[180]

This legal approach is justified based on the specificities of the spread of criminal hate speech in E2EE services. On open-ended online platforms, criminal hate speech may be directly addressed to the people targeted and immediately cause harm. Contrarily, in E2EE services, communications are confidential and shared with close contacts such as family, friends, colleagues, or collaborators. Thus content is typically shared among like-minded contacts. Such private communications among like-minded people may lead to extremism and radicalisation in places referred to as "echo chambers".[181]

Applying the legal criteria to determine which hate speech in E2EE may qualify as the most severe cases of hate speech, it is important to analyse the contextual variables (Section 4.2.3.1). Particularly relevant for criminal hate speech shared in E2EE services are: i) the content of the speech; ii) the reach and form of dissemination; iii) the nature and size of the audience; and, iv) the imminence or likelihood that the speech leads, directly or indirectly, to harmful consequences.

---

180 Grounded on international human rights law also with ICCPR, Arts. 20 and 19.

181 Ludovic Terren and Rosa Borge-Bravo (2021) Echo Chambers on Social Media: A Systematic Review of the Literature, available at <https://rcommunicationr.org/index.php/rcr/article/view/94/90> accessed 7 Sep 2023.

Assessing the first variable, hateful content shared on E2EE services may range from insults, incitement, discrimination, to incitement to violence. In the case of insults or discriminatory comments that are shared between people who are not the target of such comments, there is in itself no direct harm.[182] Nevertheless, hate speech as incitement to violence that is communicated without the knowledge of the targeted people can be an indicator of the imminence of harm, in which case it is important to assess further contextual variables applicable to E2EE services.

The second and third contextual variable can be investigated together, i.e. the reach and form of dissemination as well as the nature and size of the audience. E2EE services, with its privacy preserving features and with increasing technical affordances to create large groups around 1000 users, arguably constitute one of the most appealing digital environments for criminal activity. To recall, Signal allows for the creation of groups with around 1000 users,[183] WhatsApp of up to 5000 users,[184] and Telegram around 200,000 users.[185] This Chapter conceptualizes the corporate HRDD of internet intermediaries providing E2EE services to groups with high numbers of users. Grounding the HRDD analysis in the element of reach offers the best human right safeguard.

Fourth, all the variables examined above contribute to the analysis of the imminence or likelihood of harmful consequences on E2EE services. To summarise, a case of incitement to violence, shared with a large group of hate mongers, in a confidential and privacy preserving way such as E2EE services, represents an environment likely to lead to harmful consequences.[186]

This Chapter claims that criminal hate speech in the form of incitement to violence, targeting historically or systematically oppressed people, shared in E2EE services in large groups of like-minded people does meet the higher thresholds to be considered one of the most serious forms of hate speech. Thus, restrictions on the right to data protection (and thus on the rights to freedom of expression and association) may be implemented if abiding by the legal

182 Though proven in multiple social studies linking the prevalence of hate crimes in communities with high rates of hate speech (n 34).

183 Signal Support, Group chats, available at <https://support.signal.org/hc/en-us/articles/360007319331-Group-chats#:~:text=Admin%20controls%20of%20who%20can%20send%20messages%20and%20start%20calls,Size%20limit%20of%201000> accessed 7 Sep 2023.

184 WhatsApp Help Center, How to add and remove group participants, available at <https://faq.whatsapp.com/841426356990637/?locale=en_US&cms_platform=web&cms_id=841426356990637&draft=false> accessed 7 Sep 2023.

185 Telegram Group Chats on Telegram, available at <https://telegram.org/faq#:~:text=With%20Telegram%2C%20you%20can%20send,for%20broadcasting%20to%20unlimited%20audiences> accessed 7 Sep 2023.

186 Motafa Rachwani and Christopher Knaus (The Guardian, 2023) Videos urged counter-protesters to attack LGBTQ+ activists outside Sydney church, available at <https://www.theguardian.com/australia-news/2023/mar/22/videos-urged-counter-protesters-to-attack-lgbtq-activists-outside-sydney-church> accessed 7 Sep 2023.

requirements in Article 10(2) ECHR. Currently, the regulatory framework does not address this need to conduct a legal analysis between the right to safety and life and the right to privacy in the cases of incitement to violence in E2EE services. The following analysis seeks to address this legal loophole.

4.5.2 Procedural regulation

4.5.2.1 HRDD responsibilities of E2EE to counter incitement to violence

As examined in Section 4.3, E2EE services must comply with the HRDD framework. The corporate HRDD responsibilities of E2EE include: a policy commitment to respect human rights; the implementation of a HRDD process; remedial responsibilities; and the need to cooperate with law enforcement.

Applying the specific European HRDD standards to E2EE services, as established by the CSDDD, the policy commitment covers the responsibility to respect the Genocide Convention, ICCPR and ICERD, namely right to life and security,[187] violation of the prohibition of torture, cruel, inhuman or degrading treatment.[188] Subsequently, the HRDD process must be ongoing throughout the businesses operations and supply chain relationships and must aim to identify, prevent, mitigate, and provide for remedies for adverse impacts on human rights.

This is all the more reinforced by the European standards[189] that establish stronger HRDD responsibilities for internet intermediaries comprising higher risk to human rights. Internet intermediaries providing E2EE services can be associated with a more significant risk as the privacy-preserving setting may increase criminal activity.

Regarding the HRDD responsibility to identify adverse human rights impacts under the DSA, though there is no general monitoring obligation, internet intermediaries may be requested by national authorities to carry out specific monitoring based on national legislation or Union law.[190] As a result, there may be a basis for a request for monitoring in cases of imminent threats to the right to life Incitement to violence would meet this legal requirement.

Regarding the prevention and mitigation responsibilities stemming from HRDD, E2EE services should reflect in their terms of service the content that they do not host hate speech and state that they remove criminal hate speech. This is followed by the HRDD responsibility to, upon notice or awareness,

187 UDHR, Art. 3; ICCPR, Art. 6.
188 Article 5 UDHR, ICCPR Article 7.
189 DSA and CM/Rec(2022)16.
190 DSA, Recital 30.

remove criminal hate speech.[191] For cases that would not qualify as criminal hate speech and which would therefore require a more detailed contextual analysis, internet intermediaries should consider deamplification techniques.[192]

Furthermore, internet intermediaries, including those providing E2EE services, have the HRDD responsibility to cooperate with law enforcement if they suspect that a criminal offence involving a threat to the life or safety of a person has taken place, is taking place, or is likely to take place.[193]

4.5.2.2 Technical implementation: disruption as the minimum legal standard

This Chapter suggests the expansion of the HRDD framework to include the implementation of a minimum HRDD responsibility to disrupt large groups in E2EE services sharing incitement to violence towards historically or systematically targeted communities. This Chapter proposes a minimum HRDD responsibility broadly composed of six points which, similarly to the HRDD framework, should happen on an ongoing basis and throughout the business' operations. The possible human rights risks and suggested safeguards associated with this standard are explored in Section 4.5.3.

1) Creation of database: The legislators, in consultation with human rights organisations and civil society representing historically or systematically oppressed communities, would employ human rights standards and critical theory to create a database of minimum hateful expressions amounting to "incitement to violence". Such a database should adopt a strict interpretation of incitement to violence, guided by the expressed acknowledgement of the intersectionality of historical or systematic systems of oppression. This database must be publicly accessible. The legislators must expressly regulate the detailed requirements of the proposed HRDD standard, namely: the strict approach to the conceptualization of incitement to violence; the limited permission for process of metadata; the disruption techniques; the cooperation with law enforcement; and, the need for E2EE services to reflect these requirements in the terms of service.

2) Explain in terms of service: Internet intermediaries providing E2EE services[194] should communicate in their terms of service the database and explain the HRDD standard in their terms of service.[195] The HRDD standard would impact E2EE services allowing large size groups should explain the encryption changes in large

191 Though the CM/Rec(2022)16 suggests that any type of hate speech be removed by IS, this research disagrees with this legal approach due to the dangers of misapplication of more complex legal reasonings for hate speech cases that are not clearly criminal hate speech.

192 CM/Rec(2022)16. Deamplification is when the platform intentionally decreases the virality of certain content by adjusting their content moderation algorithms.

193 DSA, Art. 18.

194 In this section, references to internet intermediaries refer to internet intermediaries providing E2EE services and allowing large size of groups or communities.

195 DSA, Art. 14(2).

groups. In large groups, the encryption could change to homomorphic encryption and hashing to enable detecting of incitement to violence, without revealing the person's identity. Following the detection of incitement to violence as per the database, E2EE services could employ disruption techniques such as temporarily blocking the group's activity or, if systematic violations occur, the group could be broken down.

3) Monitor "the size of the audience" and "reach": Internet intermediaries have the HRDD responsibility to monitor the contextual variables of "size of the audience"[196] and "reach" deriving from human rights standards. Given the state-of-the-art concerning the messaging applications,[197] this research considers large groups the ones with over 500 users.[198] Metadata could be employed to monitor the size of the group and approximate location.[199] No additional metadata should be monitored or archived by the E2EE services. The reason to limit the monitoring of location to the city-level is because most law enforcement structures are organized from national to city-level.

4) Run homomorphic encryption or perceptual hashing: Internet intermediaries could employ homomorphic encryption to detect known[200] text, or perceptual hashing if the content combines known image and known text. This step is further detailed below.

5) Disruption techniques: Internet intermediaries could employ disruption techniques following the detection of incitement to violence in large groups. Such techniques could include freezing and, for cases of systematic breaches, dividing the group.

6) Cooperation with law enforcement: Internet intermediaries to share with law enforcement,[201] the time and approximate location of the user posting incitement to violence. A location monitored at the city level would enable already existing law enforcement structures to deploy their offline preventive criminal law enforce-

196 For a detailed analysis of the differences between scale and size in AI content moderation, see Tarleton Gillespie (2020) Content moderation, AI and the question of scale, *Big Data and Society*, available at <https://journals.sagepub.com/doi/pdf/10.1177/2053951720943234> accessed 7 Feb 2024.

197 *E.g.*, Idowu Omisola (2023) WhatsApp Community vs. WhatsApp group: What's the Difference? available at <https://www.makeuseof.com/whatsapp-community-vs-whatsapp-group-difference/> accessed 7 Sep 2023. Also, see section 5.1.

198 This number would have to be revisited based on the evolution of the size of groups in E2EE services.

199 Importantly, depending on the Internet Protocol (IP) address, metadata on location may reveal regional location but not city details. In the latter scenario, this Chapter suggests a regional approach. Monique Danao (2023) What can someone do with your IP address? available at <https://www.forbes.com/advisor/business/what-can-someone-do-with-ip-address/#:~:text=IP%20addresses%20can%20be%20used,where%20your%20device%20is%20located.> accessed 5 Feb 2024.

200 As per the database classification in point 1.

201 A possibility would be to share first with EUROPOL and INTERPOL, prior to sharing with national law enforcement bodies, as a means to attribute stronger check-and-balances in light of international human rights law.

ment mandate.[202] No extra metadata should be monitored, archived, nor shared with law enforcement bodies. Internet intermediaries to archive results of perceptual hashing technique and share such results only in the event of being solicited by criminal courts; with an emphasis on facilitating the work of the International Criminal Court (ICC) for investigative purposes of international crimes.[203]

Regarding point 4 above, we propose a high-level technical architecture that depicts how homomorphic encryption could be used to obtain a secure solution for classifying textual messages (but similarly also for images), in such a way that the server only learns the final warning flag. The client is in control of the decryption process to avoid the server learning additional information about its message.

Figure 4 below outlines this Chapter's proposal of a homomorphic approach to secure message analysis. In this setup, a E2EE client and a server collaborate in a secure manner for the analysis of the client messages. The server will never see the exact message contents, but will analyse the encrypted client messages by counting the number of forbidden words (from a known list) and comparing that number with a known threshold. The client is asked to decrypt the end result: the binary flag indicating whether a message warning should be raised. By using the technique decryption,[204] the client is also asked to deliver a mathematical proof that the decrypted flag is indeed the result of a correct decryption.[205]

This homomorphic approach can be summarised in the following steps: (1) the client sends the homomorphically encrypted message [E(M)] to the server; (2) the server counts the number of forbidden words in the message and compares the number with a threshold, in the encrypted domain, i.e. while remaining oblivious of the message contents; (3) the server produces an encrypted binary message flag; (4) the encrypted flag is sent to the client; (5) the client decrypts the flag and generates a proof of correct decryption; (6) the serves receives the flag and proof, enabling the verification of the flag.

202 To reiterate, this research recognizes that many law enforcements structures abuse their power and perpetrate historical or systematic oppressions. This Chapter is seeking to provide legal avenues capable of clarifying how law enforcement bodies can operationalize their mandate in a human rights compliant manner, which subsequently can also facilitate accountability systems for when law enforcement does not comply with the human rights framework.

203 Importantly, information should not be deleted to prevent cases such as the YouTube deletion of Syrian Archives, see Kate O'Flaherty (Wired, 2018) YouTube Keep deleting evidence of Syrian chemical weapon attacks, available at <https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video> accessed 7 Sep 2023.

204 Kristian Gjøsteen, Thomas Haines, Johannes Müller, Peter Rønne, and Tjerand Silde 'Verifiable decryption in the head' (2022) Australasian Conference on Information Security and Privacy, Springer International Publishing, 355-374.

205 In theory, the client could opt out of this decryption process, leaving some autonomy on their side.

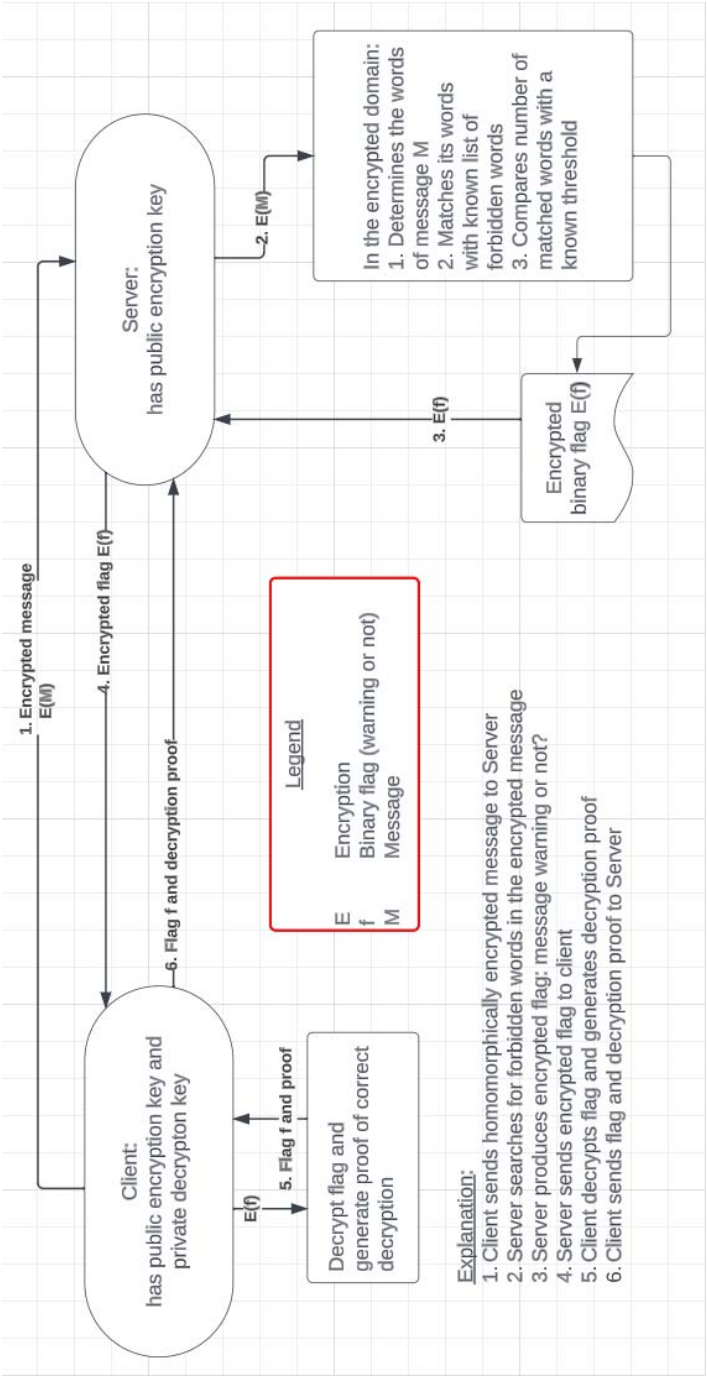

Figure 4 – Homomorphic approach to secure message analysis.

4.5.2.3 Legal implementation

This HRDD standard could have legal grounding in Article 9 of the DSA which establishes the possibility for internet intermediaries to receive orders from national judicial or administrative authorities, on the basis of *inter alia* European Union Law or national law in compliance with Union law. On the one hand, Union Law may soon impose standardised obligations on EU member states to protect their citizens from hate speech should hate speech become part of the EU crimes.[206] On the other hand, national law in EU member states already establishes the right to life and safety. As a result, under this basis, the proposed HRDD responsibility to monitor incitement to violence on large groups operating in E2EE services could already be implemented. This aligns with the lawful basis under data protection law as per Article 6 of the GDPR.[207]

The element of cooperation with law enforcement finds legal grounding in Article 18 of the DSA, which articulates that internet intermediaries shall promptly inform law enforcement if they become aware of information giving rise to suspicion that a criminal offence involving a threat to the life or safety of a person or multiple people.

The proposed HRDD standard is both a HRDD measure and a high-risk Artificial Intelligence system in the context of the Artificial Intelligence Act.[208] This HRDD standard would be considered high-risk because it would be an AI system "intended to be used in support of law enforcement authorities on behalf of law enforcement authorities to assess the risk of a natural person to become a victim of criminal offences.[209] As a result, E2EE services implementing this standard would have to comply with stricter human rights responsibilities as per the AI Act.[210]

4.5.3 Critical analysis: human rights safeguards

This section provides a critical analysis concerning the human rights safeguards in the proposed HRDD standard by expanding on the compliance with the legal frameworks related to the rights to freedom of expression, to freedom of assembly and association, and to data protection.

The compliance of the proposed HRDD standard with the human rights provisions on freedom of expression and freedom of association can be inter-

206 See supra (n 30).
207 See Section 2.3.3.
208 This overlap between HRDD standards and AI systems potentially considered high-risk under the AI Act is likely to increase as businesses develop AI methods to monitor the compliance of their services with human rights.
209 AI Act, Annex III, Article 6(b).
210 AI Act, Chapter 3.

preted together as they are accompanied by the same legal requirements for any eventual restriction. To clarify, the proposed standard complies with Articles 10 and 11 of the ECHR because it would be prescribed by law (Section 4.5.2.3), in pursuit of public safety, and it would be addressing a pressing social need that is the prevention of hate crimes.

Furthermore, the proposed HRDD standard is proportional in that it is the least intrusive measure for three main reasons. First, the proposed HRDD standard would follow a strict conceptualization of incitement to violence based on intersectionality of historical or systematic systems oppression. Additionally, the incitement to violence database would have to be translated into all languages currently used in online platforms[211] The translation should be done through community classification of incitement to violence with the support of human rights scholars, practitioners, or targeted communities. The database would have to be publicly communicated in the terms of service.[212] The incitement to violence database, without the context for the incitement to violence, can detect cases where the speaker is a person reporting a case of incitement to violence.[213] This Chapter suggests the exploration of certified accounts for human rights activists[214] and the automatic sharing of helplines for human rights activists.

Second, the proposed HRDD standard would be the least intrusive technical solution because it would require the regulation of collection of metadata, of privacy preserving detection methods, of the disruption techniques, and of the cooperation framework with law enforcement. The standard proposed is that, aside from metadata on the group size and approximate location, no other metadata should be collected by E2EE services. Additionally, the proposed standard guarantees the users' privacy because it relies on homomorphic encryption and hashing techniques. Furthermore, the disruption techniques employed are likewise the least intrusive possible as information detected should not be deleted.[215] The suggested disruption techniques would

211 The translation costs would be supported by the platforms providing E2EE services.

212 *E.g.*, in Europe the European Observatory of Online Hate (EOOH), could assist also in this task too should it ensure representativeness from targeted groups.

213 For example, someone calling for help and reproducing the attack message of the perpetrator. Such content would potentially also be picked up in such a digital intervention.

214 Notably, the possibility for the restriction on the right to freedom of assembly and association also applies to governments and law enforcement bodies. Civil or military servants are not to be conflated with human rights activists. This is all the more important given the growing infiltration of violent extremism in law enforcement bodies. *E.g.* Hassan Kanu (Reuters, 2022) Prevalence of white supremacists in law enforcement demands drastic change, available at <https://www.reuters.com/legal/government/prevalence-white-supremacists-law-enforcement-demands-drastic-change-2022-05-12/> accessed 7 Feb 2024.

215 Contrarily to CSAM, which if posted causes immediate harm and thus requires a more difficult balance between the removal and the non-removal, incitement to violence in E2EE services does not cause immediate harm and thus an intervention would not necessarily involve removal of content. Disruption techniques not including removal would be less intrusive on freedom of expression than other previous proposals to counter illegal content

prioritize freezing over division of the group. Division of the group would only occur after systematic breaches of the HRDD standard and recurrent detection of incitement to violence.

Third, the proposed HRDD standard would comply with transparency requirements. A timeframe would have to be established to explain to users the new HRDD standard. Internet intermediaries to submit to the DSA Coordinator annual reports on the implementation of the proposed HRDD standard.

The compliance of the proposed standard with the human rights provisions on data protection under Articles 5 and 6 of the GDPR and Article 5 of the e-Privacy Directive for the following reasons. First, it would have a lawful basis (Section 4.5.2.3). Second, it would be shared beforehand with users through the terms of service and through a specific notification in E2EE groups over the minimum threshold alerting that, in such large groups, it is not permitted to share incitement to violence according to the database in the terms of service. Third, users in large groups would therefore be informed and would give their consent to the application of this standard which would be carried out in the public interest of protecting the right to safety and life of people historically or systematically targeted by hate speech.

In effect, this would be a detection order regime but, contrary to previously proposed detection order regimes in the case of CSAM and terrorism, this has a narrower and more concrete scope with clear human rights safeguards outlined. Table 2 below summarises the proposed HRDD standard.

This Chapter acknowledges that the standard hereby proposed alone will not end incitement to violence on E2EE services for various reasons. For instance, language can be coded to avoid matching that in the database, the group size can likewise be circumvented easily, and there are a multitude of alternative online services used to spread incitement to violence.[216] Nevertheless, the standard proposed in this Chapter serves a key purpose – it clarifies the corporate human rights responsibilities of E2EE services by reiterating the prohibition of incitement to violence in human rights law. Consequently, it is expected to contribute to the deterrence objective of regulatory framework, decrease incitement to violence on E2EE services, and subsequently decrease offline hate crimes.

on E2EE services. See Section 3.3.

216 See *e.g.*, Andrew D. Murray (2011) Nodes and gravity in Virtual Space, 208, *Legisprudence,* 10.5235/175214611797885684.

Table 2 – Summary of proposed HRDD standard

Phase	*Actor*	*Method*	*Action*	*Human rights safeguards*
1	Legislators in consultation with human rights organizations and civil society	Human rights standards	Create database of "incitement to violence"	- strict linguistic interpretation - intersectional - historical or systematic oppression - in languages currently spoken on online platforms
2	Internet intermediaries E2EE, only the ones enabling groups over 500 users	Human rights standards	Explain in terms of service	- legal clarity and foreseeability - users' consent
3	Internet intermediaries E2EE, only the ones enabling groups over 500 users	Metadata	Monitor "the size of the audience" and "reach"	- application of contextual variables used to identify the most serious forms of hate speech
4	Internet intermediaries E2EE, only the ones enabling groups over 500 users	Homomorphic encryption	Run homomorphic encryption or perception hashing if the content combines image and text, ex post monitoring	- users' privacy is guaranteed
5	Internet intermediaries E2EE, only the ones enabling groups over 500 users	Homomorphic encryption	Disruption techniques (showing support helplines, freezing groups, dividing groups)	- post is not deleted, thus freedom of expression is not disproportionally compromised - users' privacy is guaranteed - the possibility for the restriction on the right to freedom of assembly and association also applies to governments and law enforcement bodies posting incitement to violence.

Phase	*Actor*	*Method*	*Action*	*Human rights safeguards*
6	Internet intermediaries E2EE, only the ones enabling groups over 500 users, to cooperate with law enforcement	International cooperation	Cooperation with law enforcement (sharing approx. time and location of user to support law enforcement monitor incitement to violence in public settings)	- could identify target groups and share information and location with governments so that more law enforcement would be deployed to protect historically marginalized communities. However, studies show records of law enforcement abusing their power and being the perpetrators of human rights violations of the targeted groups. A strict monitoring of the law enforcement activities would be essential.

4.6 Conclusion

This research tackles the pressing problem of having digital spaces accessible to large numbers of users (some reaching the thousands all at once), prone to the rise of criminal activity, and with no accountability. As one of the consequences, people targeted by hate speech are now at a higher risk and with less protection mechanisms provided by democratic law enforcement bodies. At the same time, such digital spaces offer essential secure and confidential communication for human rights activist.

The human rights framework is trying to adjust and HRDD standards have been proposed in the field of CSAM and terrorism. However, these legal strategies hinder human rights provisions on freedom of expression, freedom of association, privacy, or data protection.

This Chapter applies interdisciplinary methods comprising human rights, digital technologies, and international cooperation to propose an innovative and proportional legal interpretation of technological developments expanding the HRDD of online platforms, and especially of very large online platforms, providing E2EE services in the European context to not host criminal hate speech in the form of incitement to violence. The HRRD standard complies with freedom of expression, association, and data protection as it founded on disruption techniques applicable only to groups over 500 users. Such disruption techniques encompass, freezing, or in worst case scenarios, dividing groups. Finally, to ensure the protection of human rights activists, the HRDD standard proposes automatically showing helpline numbers and creating certified E2EE accounts for human rights activists to denounce human rights violations. Moreover, this Chapter is innovative in the proposal for regulation of metadata in E2EE services in a manner compliant with the GDPR and with the e-Privacy Directive by suggesting that only time and approximate location

be collected and made available to law enforcement. E2EE services are required to archive data inciting to violence for potential use in international criminal actions.

This Chapter proposes a minimum HRDD standard, based on homomorphic encryption, to counter incitement to violence, legally classified as within the most serious cases of hate speech, in E2EE services provided by online platforms. Very large online platforms would have the heightened responsibility to adhere to this HRDD standard. The HRRD differs from the corporate liability framework, which would still have to be developed in future research and encompasses different considerations in terms of which legal incentives of penalties to introduce, that is outside the scope of this Chapter. Additionally, future research is needed on the monetization of E2EE services and on the introduction of features such as self-destructing messages.

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

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

ABSTRACT

Online platforms have adopted business models enabling the proliferation of hate speech. In some extreme cases, platforms are being investigated for employing algorithms that amplify criminal hate speech such as incitement to genocide. Legislators have developed binding legal frameworks clarifying the human rights due diligence and liability regimes of these platforms to identify and prevent hate speech. Some of the key legal instruments at the European Union level include the Digital Services Act, the proposed Corporate Sustainability Due Diligence Directive, and the Artificial Intelligence Act. However, these legal frameworks fail to clarify the remedial responsibilities of online platforms to redress people harmed by criminal hate speech caused or contributed to by the platforms. This Chapter addresses this legal vacuum by proposing a comprehensive remedial responsibilities framework for online platforms which caused or contributed to criminal hate speech based on the general corporate human rights responsibilities framework.

5.1 INTRODUCTION

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

---

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

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

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

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

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

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

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

5 Amnesty International, 'Myanmar: The social atrocity: Meta and the right to remedy for the Rohingya' (2022) https://www.amnesty.org/en/documents/ASA16/5933/2022/en/ (accessed 28 May 2024); Human Rights Council, 'Report of the independent international fact-finding mission on Myanmar' (2018) A/HRC/39/64, https://www.ohchr.org/en/press-releases/2018/09/myanmar-un-fact-finding-mission-releases-its-full-account-massive-violations?LangID=E&NewsID=23575 (accessed 28 May 2024), Para. 74.

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

7 UN Human Rights Council, 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie' (2011) A/HRC/17/31 (UNGPs).

8 UNGPs (note 7), Principle 15.

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

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

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

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

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

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

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

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

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

14 DSA, note 11, Art. 41.

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2 Criminal hate speech on online platforms

5.2.1 European standards on criminal hate speech

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2.2 The role of online platforms

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

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

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

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

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

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

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

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

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

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

---

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

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

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

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

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

44 Tarleton, note 39.

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

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

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

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

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

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

48 Michel Rosenfeld, Hate Speech in Constitutional Jurisprudence: A Comparative Analysis Conference: The Inaugural Conference of the Floersheimer Center for Constitutional Democracy: Fundamentalisms, Equalities, and the Challenge to Tolerance in a Post-9/11 Environment, 24 Cardozo L. Rev. 1523, 1565 (2002). CM/Rec(2022)16, , note 17, Explanatory memo, Para. 32.

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

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

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

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

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

54 Budinova and Chaprazov v. Bulgaria Para. 63.

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

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

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

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

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

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

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

59 Amnesty International, note 5, 42.

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

61 Amnesty International, note 5, 42.

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

63 Amnesty International, note 5.

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

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

5.3 Right to remedy for criminal hate speech online

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

5.3.1 Harm caused by hate speech

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

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

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

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

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

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

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

5.3.2 State's duty to ensure access to remedy

5.3.2.1 European standards on remedies

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3.2.2 Remedies for gross human rights violations

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

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

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

96 United Nations, Resolution adopted by the General Assembly on 16 December 2005, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, A/RES/60/147, II(3).

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

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

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

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

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

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

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

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

5.4 General framework: corporate remedial responsibilities for online platforms

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

5.4.1 Modes of corporate responsibility

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

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

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

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

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

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

---

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

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

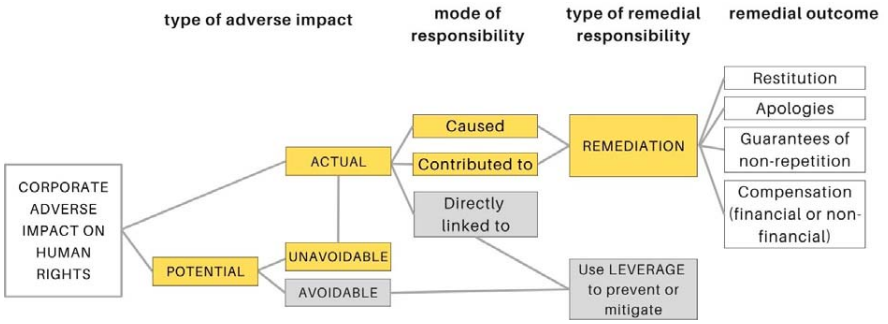

Figure 5 – Corporate remedial responsibilities for adverse human rights impacts

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

5.4.2 Remedial processes

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

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

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

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

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

5.4.3 Remedial outcomes

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

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

---

113 UNGPs Guide, note 102, Q. 66.

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

115 Eva Nave and Lottie Lane, note 42.

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

117 UNGPs, note 7, Principle 20.

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

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

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

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

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

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

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

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

5.5.1 Challenges with current framework

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

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

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

5.5.1.1 Corporate remedial responsibilities in the EU

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

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

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

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

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

5.5.1.2 Remedial responsibilities in the Digital Services Act

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

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

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

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

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

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

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

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

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

---

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

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

141 DSA, note 11, Recital 59.

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

143 DSA, note 11, Art. 21.

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

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

5.5.1.3 Complementary corporate remedial frameworks for hate speech online

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

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

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

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

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

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

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

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

5.5.2 Proposed standards for a comprehensive framework

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

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

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

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

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

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

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

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

5.5.2.1 Restitution and satisfaction as amplification of survivors' speech

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

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

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

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

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

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

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

5.5.2.2 Compensation and rehabilitation beyond area of services

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

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

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

---

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

166 European Union (note 99).

5.5.2.3 Guarantees of non-repetition as business models' change

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

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

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

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

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

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

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

5.6 Conclusion

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

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

173 Gelber (note 30), 401.

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

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

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

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

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

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

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

6 Conclusion[1]

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

6.1 FINDINGS RELATED TO THE PROBLEM STATEMENT AND RESEARCH QUESTIONS

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

> "Building on a critical conceptualization of online hate speech, and more specifically on criminal hate speech, deriving from the European regulatory and policy framework, how can European legislators, both at the European Union and at the Council of Europe levels, clarify the responsibilities of online platforms to counter online hate speech whilst upholding fundamental rights?"

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

---

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

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

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

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

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

6.1.1 Legal conceptualization of hate speech

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

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

4 DSA, Art. 43.

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

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

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

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

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

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

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

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

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

11 *E.g.* TikTok, WeChat.

12 *E.g.* Telegram.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

35 Center for Democracy &Technology (n 34).

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

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

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

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

36 European Union, CSDDD, Annex I.

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

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

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

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

38 European Union, DSA, Recital 14.

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

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

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

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

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

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

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

43 Amnesty International (n 42), 43.

44 Amnesty International (n 42), 66.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.2 RECOMMENDATIONS

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

6.3 Areas for future research

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Summary

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

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

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

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

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

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

Chapter 3 advances specific preventive HRDD responsibilities applicable to online platforms countering online hate speech. Increased attention is being paid to the corporate HRDD responsibilities applicable to online platforms to counter online hate speech. At the European Union level, cross-sector initiatives regulate the rights of marginalised groups and establish HRDD responsibilities for online platforms to expeditiously identify, prevent, mitigate, remedy and remove online hate speech. Nevertheless, the HRDD framework

applicable to online hate speech has focused mostly on the platforms' responsibilities throughout the course of their operations – guidance regarding HRDD requirements concerning the regulation of hate speech in the platforms' Terms of Service is missing. This Chapter employs a conceptualisation of criminal hate speech as explained in the Council of Europe Committee of Ministers' Recommendation CM/Rec(2022)16, Paragraph 11, to develop specific HRDD responsibilities. This research includes an empirical qualitative analysis of three case studies: Facebook (Meta Platforms, Inc.), X Corp. (previously Twitter, Inc.), and YouTube. This empirical analysis assesses the compliance of the platforms' Terms of Service with the conceptualisation of criminal hate speech in CM/Rec(2022)16. This Chapter claims that online platforms should, as part of emerging preventive HRDD responsibilities within Europe, respect the rights of historically oppressed communities by aligning their Terms of Service with the conceptualisation of criminal hate speech in European human rights standards.

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

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

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

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

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

Samenvatting (Dutch Summary)

De bestrijding van online haat vanuit fundamenteelrechtelijk perspectief

Dit manuscript onderzoekt juridische benaderingen die in overeenstemming zijn met de mensenrechten om haatzaaiende uitlatingen op online platforms tegen te gaan. Er is een groeiende prevalentie van haatzaaiende uitlatingen op online platforms. Online platforms hebben zelfregulerende beleidsmaatregelen ontwikkeld om haatzaaiende uitlatingen tegen te gaan. Echter, dergelijke private regelgevende kaders blijven vaak ondoorzichtig en ontberen democratische handhavings- en herstelmechanismen. Dit onderzoek bouwt voort op een kritische conceptualisering van online haatzaaiende uitlatingen, gebaseerd op het Europese regelgevende en beleidskader, om juridische wegen te onderzoeken en voor te stellen die de zorgvuldigheidsplichten met betrekking tot mensenrechten (HRDD) van online platforms versterken. Dit met als doel om strafbare online haatzaaiende uitlatingen tegen te gaan en tegelijkertijd fundamentele rechten te waarborgen.

Hoofdstuk 1 legt de basis door de context en sociale relevantie te presenteren, de probleemstelling en onderzoeksvragen te introduceren en de methodologie en reikwijdte uit te leggen. Online social media-platforms kunnen in brede zin worden omschreven als internet-hostingdiensten die door gebruikers gegenereerde inhoud opslaan, beoordelen, promoten of degraderen voor het algemene publiek via groepen, gepersonaliseerde nieuwsfeeds en berichtenapplicaties. Met meer dan de helft van de wereldbevolking als actieve gebruikers van online socialemediaplatforms, en met het gebruik van deze online omgevingen om fundamentele mensenrechten uit te oefenen, zoals vrijheid van meningsuiting, vrijheid van vergadering en vereniging, heeft de verspreide inhoud en de implementatie van moderatiebeleid een steeds grotere impact op wereldschaal. Mensenrechtenactivisten, klokkenluiders, voormalige werknemers van online platforms en zelfs de Verenigde Naties hebben gewaarschuwd dat sommige sociale mediabedrijven met hun bedrijfsmodellen niet alleen hebben gefaald in het verwijderen van haatzaaiende uitlatingen, maar deze zelfs hebben versterkt. Verdere studies waarschuwen ook dat de toename van haatzaaiende uitlatingen in digitale omgevingen kan resulteren in offline haatzaaiende uitlatingen en haatmisdrijven.

Als reactie op deze ontwikkelingen en onder druk van staten, mensenrechtenactivisten en het maatschappelijk middenveld, zijn sommige online plat-

forms begonnen met zelfregulering van haatzaaiende uitlatingen, het delen van gegevens over de prevalentie van haatzaaiende uitlatingen en het opzetten van toezichtsorganen voor beroepsprocedures inzake contentmoderatie. Echter, dergelijke zelfregulerende inspanningen worden vaak bekritiseerd vanwege het niet naleven van mensenrechtennormen. Belangrijke kritiekpunten zijn onder meer: i) de conceptualisering van haatzaaiende uitlatingen die door platforms wordt gehanteerd; ii) de handhavingsmechanismen voor contentmoderatiebeleid; iii) de herstelmechanismen die gebruikers ter beschikking staan om contentmoderatiebeslissingen aan te vechten. In een poging om de reguleringskaders die door bedrijven worden ingesteld om online haatzaaiende uitlatingen tegen te gaan, democratisch te reguleren en te controleren, produceren zowel staten als internationale en regionale organisaties sectorspecifieke juridische en beleidsinstrumenten. Niettemin rijzen er discussies over de effectiviteit en geschiktheid van dergelijke regelgevende kaders bij het bevorderen van respect voor de mensenrechten van mensen die het doelwit zijn van haatzaaiende uitlatingen.

Daarom stelt dit manuscript de volgende vraag: Gebaseerd op een kritische conceptualisering van online haatzaaiende uitlatingen, en meer specifiek van strafbare haatzaaiende uitlatingen binnen het Europese regelgevende en beleidskader, hoe kunnen Europese wetgevers, zowel op het niveau van de Europese Unie als de Raad van Europa, de zorgvuldigheidsplichten met betrekking tot mensenrechten van online platforms verduidelijken om online haatzaaiende uitlatingen tegen te gaan en tegelijkertijd fundamentele rechten te waarborgen? De gebruikte methodologie is drieledig: doctrinair; vergelijkend; en interdisciplinair. Doctrinair juridisch onderzoek van juridische en beleidskaders die van toepassing zijn op online haatzaaierij probeert bestaande juridische normen, mazen en mogelijke toekomstige juridische wegen te verduidelijken. Vergelijkende juridische analyse wordt gebruikt om de afstemming, of het gebrek daaraan, tussen de normen die door onlineplatforms worden aangenomen via hun servicevoorwaarden en Europese mensenrechtennormen te onderzoeken. Interdisciplinair onderzoek combineert bevindingen uit juridische, sociologische en digitale technologiestudies om zorgen te systematiseren en contentmoderatiepraktijken voor te stellen die geschikt zijn om online haatzaaierij tegen te gaan en die voldoen aan de mensenrechten.

Hoofdstuk 2 stelt een nieuwe juridische conceptualisering van haatzaaiende uitlatingen in de Europese context voor. Huidige kaders missen een gestandaardiseerde benadering van de conceptualisering van haatzaaiende uitlatingen. Sommige conceptualiseringen zijn te breed, en andere zijn onvoldoende inclusief; te breed omdat ze leiden tot het verwijderen van legale content (bijv. verwijderingstools die legale content verwijderen die is geplaatst door gemarginaliseerde gemeenschappen), en onvoldoende inclusief omdat de context van posts door taalkundige minderheden vaak wordt genegeerd. Dit hoofdstuk analyseert het Europese regelgevingskader door de lens van de eerste juridische conceptualiseringen van haatzaaiende uitlatingen die voortvloeien uit de

kritische (rassen)theorie en (zwarte) feministische intersectionaliteitstheorie. Er zijn twee belangrijke bevindingen uit dit hoofdstuk. Ten eerste suggereert dit hoofdstuk dat het Europese regelgevingskader expliciet de conceptualisering van haatzaaiende uitlatingen door kritische rechtsgeleerden moet erkennen als uitingen die bedoeld zijn om historische of systematische onderdrukking in stand te houden. Ten tweede bepleit dit hoofdstuk dat de conceptualisering van haatzaaiende uitlatingen in de Europese context alleen juridische samenhang kan bereiken wanneer alle Europese regelgevingsinstrumenten uitdrukkelijk rekening houden met de intersectionaliteit van systemen van onderdrukking.

Hoofdstuk 3 bevordert specifieke preventieve HRDD-verantwoordelijkheden die van toepassing zijn op onlineplatforms die online haatzaaiende uitlatingen tegengaan. Er wordt meer aandacht besteed aan de HRDD-verantwoordelijkheden van bedrijven die van toepassing zijn op onlineplatforms om online haatzaaiende uitlatingen tegen te gaan. Op het niveau van de Europese Unie reguleren sectoroverschrijdende initiatieven de rechten van gemarginaliseerde groepen en stellen HRDD-verantwoordelijkheden vast voor onlineplatforms om online haatzaaiende uitlatingen snel te identificeren, voorkomen, beperken, verhelpen en verwijderen. Niettemin heeft het HRDD-kader dat van toepassing is op online haatzaaiende uitlatingen zich vooral gericht op de verantwoordelijkheden van de platforms gedurende hun hele bedrijfsvoering – richtlijnen met betrekking tot HRDD-vereisten met betrekking tot de regulering van haatzaaiende uitlatingen in de Servicevoorwaarden van de platforms ontbreken. Dit hoofdstuk gebruikt een conceptualisering van criminele haatzaaiende uitlatingen zoals uitgelegd in de Aanbeveling CM/Rec(2022)16, paragraaf 11, van het Comité van Ministers van de Raad van Europa om specifieke HRDD-verantwoordelijkheden te ontwikkelen. Dit onderzoek omvat een empirische kwalitatieve analyse van drie casestudies: Facebook (Meta Platforms, Inc.), X Corp. (voorheen Twitter, Inc.) en YouTube. Deze empirische analyse beoordeelt de naleving van de Servicevoorwaarden van de platforms met de conceptualisering van criminele haatzaaiende uitlatingen in CM/Rec(2022)16. Dit hoofdstuk beweert dat onlineplatforms, als onderdeel van de opkomende preventieve HRDD-verantwoordelijkheden binnen Europa, de rechten van historisch onderdrukte gemeenschappen moeten respecteren door hun Servicevoorwaarden af ??te stemmen op de conceptualisering van criminele haatzaaiende uitlatingen in Europese mensenrechtennormen.

Hoofdstuk 4 stelt een nieuwe wettelijke minimumstandaard voor die de verantwoordelijkheden van onlineplatforms die E2EE-diensten aanbieden op het gebied van mensenrechten uitbreidt om een ??categorie van criminele haatzaaiende uitlatingen te beperken: aanzetten tot geweld. Diensten die door onlineplatforms worden aangenomen, hebben de verspreiding van online haatzaaiende uitlatingen mogelijk gemaakt. Met name end-to-end gecodeerde (E2EE) diensten staan ??onder toenemende controle voor het hosten van haatzaaiers. Juristen en wetshandhavers worstelen met het conceptualiseren

van de verantwoordelijkheden van E2EE-diensten om geen haatzaaiende uitlatingen te hosten zonder de rechten van gebruikers op vrijheid van meningsuiting, vereniging, privacy of gegevensbescherming onevenredig te beïnvloeden. Na het vaststellen van het algemene HRDD-kader voor HRDD van bedrijven met kunstmatige intelligentie om criminele haatzaaiende uitlatingen te beperken, gaat dit hoofdstuk dieper in op de digitale technologieën en encryptiefuncties die worden gebruikt voor contentmoderatie in E2EE-diensten. Deze analyse past het HRDD-kader toe in combinatie met homomorfe encryptie, metadata en hashing op geselecteerde criminele haatzaaiende uitlatingen die aanzetten tot geweld. Bovendien verduidelijkt dit hoofdstuk de normen voor samenwerking tussen onlineplatforms en wetshandhaving in de context van aanzetten tot geweld in grote groepschats op E2EE-services. Tot slot stelt hoofdstuk 4 een nieuwe wettelijke norm voor die de corporate HRDD van onlineplatforms die E2EE-services aanbieden uitbreidt door middel van regulering en toepassing van metadata, hashing en homomorfe encryptie om aanzetten tot geweld in grote groepen op E2EE-services te verstoren.

Hoofdstuk 5 stelt een uitgebreid kader voor herstelverantwoordelijkheden voor onlineplatforms voor die criminele haatzaaiende uitlatingen hebben veroorzaakt of ertoe hebben bijgedragen, op basis van het algemene kader voor verantwoordelijkheden van bedrijven voor mensenrechten. Wetgevers hebben bindende juridische kaders ontwikkeld die de due diligence- en aansprakelijkheidsregimes voor mensenrechten van deze platforms verduidelijken om haatzaaiende uitlatingen te identificeren en te voorkomen. Deze juridische kaders verduidelijken echter niet de herstelverantwoordelijkheden van onlineplatforms om mensen te herstellen die schade hebben geleden door criminele haatzaaiende uitlatingen die door de platforms zijn veroorzaakt of waaraan deze hebben bijgedragen. De bijdrage van Meta aan de genocide op de Rohingya in Myanmar wordt geanalyseerd als een van de meest grondig gedocumenteerde gevallen die de maatschappelijke impact laten zien van de verantwoordelijkheden van bedrijven voor mensenrechten van zeer grote onlineplatforms die bijdragen aan de versterking van criminele haatzaaiende uitlatingen.

Dit hoofdstuk onderzoekt de toepassing van het recht op een doeltreffende remedie op gevallen van online haatzaaiende uitlatingen. Dit onderzoek onderzoekt ook de internationale normen voor het recht op remedie voor gevallen van grove schendingen van mensenrechten, waarbij wordt erkend dat sommige elementen van criminele haatzaaiende uitlatingen kunnen worden geclassificeerd als grove schendingen van mensenrechten. Na het verduidelijken van het algemene kader voor corrigerende verantwoordelijkheid van bedrijven, dat betrekking heeft op verantwoordelijkheidswijzen, herstelprocessen en herstelresultaten, verduidelijkt dit hoofdstuk dat het herstelkader van toepassing is op onlineplatforms die criminele haatzaaiende uitlatingen hebben veroorzaakt of daaraan hebben bijgedragen. Dit hoofdstuk benadrukt de noodzaak van en stelt een kader voor corrigerende verantwoordelijkheden van bedrijven op EU-niveau voor, ook voor onlineplatforms die criminele

haatzaaiende uitlatingen hebben veroorzaakt of daaraan hebben bijgedragen. Het voorgestelde kader onderzoekt garanties van niet-herhaling, restitutie en compensatie als geschikte herstelresultaten.

Hoofdstuk 6 presenteert de belangrijkste bevindingen met betrekking tot de probleemstelling en onderzoeksvragen die deze thesis motiveren, doet aanbevelingen en bespreekt gebieden voor toekomstig onderzoek. Dit hoofdstuk doet een uitgebreide reeks aanbevelingen om de verantwoordelijkheden van onlineplatforms op het gebied van mensenrechten van bedrijven te versterken om criminele haatzaaiende uitlatingen tegen te gaan. Deze aanbevelingen zijn gericht aan drie actoren, namelijk wetgevers en beleidsmakers, wetshandhavingsinstanties en onlineplatforms. Over het algemeen kunnen wetgevers en beleidsmakers met meer vertrouwen vertrouwen op het algemene HRDD-kader om preventieve, verzachtende en herstellende verantwoordelijkheden voor onlineplatforms te conceptualiseren om criminele haatzaaiende uitlatingen tegen te gaan. Wetshandhavingsinstanties moeten de oprichting van rapportage- en onderzoekskanalen voor criminele haatzaaiende uitlatingen op onlineplatforms vergemakkelijken. Onlineplatforms moeten zich houden aan het HRDD-kader en duidelijke en transparante processen ontwikkelen om criminele haatzaaiende uitlatingen die via hun diensten worden verspreid, te voorkomen, te verzachten en te herstellen. De regulering van onlineplatforms brengt complexe juridische kwesties met zich mee in verschillende onderzoeksdisciplines, met gevolgen voor een veelheid aan nationale en internationale rechtsgebieden en met betrekking tot talrijke belanghebbenden, waaronder onlineplatforms, overheidsinstanties en individuen. Gezien de voortdurend veranderende aard van de diensten die onlineplatformen aanbieden, is het van belang dat toekomstig onderzoek naar maatregelen om online haatzaaierij tegen te gaan, sterker interdisciplinair onderzoek vereist dat zich richt op mensenrechten.

Bibliography

Legal Sources

United Nations

- United Nations, General Assembly, Universal Declaration of Human Rights, 217 A (III), 10 December 1948 (UDHR).
- United Nations, General Assembly, International Convention on the Elimination of All Forms of Racial Discrimination, Treaty Series, vol. 660, 21 December 1965 (ICERD).
- United Nations, General Assembly, International Covenant on Civil and Political Rights, Treaty Series, vol. 999, p. 171, 16 December 1966 (ICCPR).

Council of Europe

- Council of Europe, Statute of the Council of Europe, May 5, 1949, 87 U.N.T.S. 103.
- Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, ETS 5, 4 November 1950 (ECHR).
- Council of Europe, Framework Convention for the Protection of National Minorities and Explanatory Report, European Treaty Series – No. 157, Doc. H9510 (1995).
- Council of Europe, Protocol 12 to the European Convention on Human Rights, Nov. 4, 2000, E.T.S. 177.
- 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, Jan. 28, 2003, E.T.S. 189.
- Council of Europe, Convention on Preventing and Combating Violence Against Women and Domestic Violence, May 11, 2011, E.T.S. 210.

European Union Primary Law

- European Union, Charter of Fundamental Rights of the European Union, OJ C 326, 26.10.2012, p. 391–407 (CFREU).
- European Union, Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union, Protocols Annexes to the Treaty on the Functioning of the European Union Declarations annexed to the Final Act of the Intergovernmental Conference which adopted the Treaty of Lisbon, signed on 13 December 2007 Tables of equivalences, OJ C 202, 7.6.2016, p. 1–388.

European Union Secondary Law

Regulations of the European Parliament and of the Council

- European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88.
- European Union, Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 Apr 2021 on addressing the dissemination of terrorist content online, OJ L 172, 17.5.2021, p. 79–109.
- 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 (DSA OJ L 277, 27.10.2022, p. 1–102.
- European Union, 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, AI Act), OJ L, 2024/1689, 12.7.2024.

Directives of the European Parliament and of the Council

- European Union, Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications), OJ L 201, 31.7.2002, p. 37–47.
- European Union, 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, 15.4.2010, p. 1–24 (AVMSD).
- European Union, Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA, OJ L 315, 14.11.2012, p. 57–73 (Victims Directive).
- European Union, Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA, OJ L 119, 4.5.2016, p. 89–131.
- European Union, 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, OJ L, 2024/1760, 5.7.2024 (CSDDD).

Council of the European Union

- European Union, 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, OJ L 328, 6.12.2008, p. 55–58.
- European Union, Council Decision (CFSP) 2020/1999 of December 2020 concerning restrictive measures against serious human rights violations and abuses, OJ L 410I, 7.12.2020, p. 13–19.
- European Union, 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).

European Parliament

- European Union, 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).
- European Union, European Parliament, Resolution of 18 January 2024 on extending the list of EU crimes to hate speech and hate crime (2023/2068(INI)), OJ C, C/2024/5733, 17.10.2024.
- European Union, 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.

European Commission

- European Union, European Commission, Code of Conduct to counter illegal hate speech online, 2016.
- European Union, European Commission, Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, OJ L 63, 6.3.2018, p. 50–61.
- European Union, European Commission, Communication From The Commission To The European Parliament And The Council A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime, COM/2021/777 final.
- European Union, European Commission, Proposal for a Directive of the European Parliament and of the Council on Corporate Sustainability Due Diligence and amending Directive (EU) 2019/1937, COM/2022/71 final.
- European Union, European Commission, Proposal for a Directive of the European Parliament and the Council on combating violence against women and domestic violence, COM/2022/105 final.

· European Union, European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse, COM/2022/209 final.
· European Union, European Commission, Proposal for a Directive of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse and replacing Council Framework Decision 2004/68/JHA (recast), COM/2024/60 final.

Other Relevant Legal Sources

· United Kingdom, Parliament of the United Kingdom, Online Safety Act (2023), c. 50, 26 October 2023, available at <https://www.legislation.gov.uk/ukpga/2023/50/section/12/enacted> accessed 11 August 2024.
· Germany, Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act), Ministry of Justice and Consumer Protection, 12 July 2017, English Version, available at: <https://www.bmj.de/SharedDocs/Downloads/DE/Gesetzgebung/RefE/NetzDG_engl.pdf?__blob=publicationFile&> accessed 12 January 2025.

Case Law

European Court of Human Rights (ECtHR)[1]

Judgements, decisions and advisory opinions of the "new" Court (as from 1 November 1998)

Alekhina and Others v. Russia, no. 22519/02, 13 April 2006, ECLI:CE:ECHR:2018:0717JUD003800412.
Association ACCEPT and Others v. Romania, no. 19237/16, 1 June 2021, ECLI:CE:ECHR:2021:0601JUD001923716.
Beizaras and Levickas v. Lithuania, no. 41288/15, 14 January 2020, ECLI:CE:ECHR:2020:0114JUD004128815.
Belkacem v. Belgium, no. 343667/14, 27 June 2017, ECLI:CE:ECHR:2017:0627DEC003436714.
Budayeva and Others v. Russia, nos. 15339/02 and 4 others, ECHR 2008 (extracts), ECLI:CE:ECHR:2008:0320JUD001533902.
Ceylan v. Turkey [GC], no. 23556/94, ECHR 1999-IV, ECLI:CE:ECHR:1999:0708JUD002355694.
Delfi AS v. Estonia [GC], no. 64569/09, ECHR 2015, ECLI:CE:ECHR:2015:0616JUD006456909.
Dicle v. Turkey (no. 2), no. 46733/99, 11 April 2006, ECLI:CE:ECHR:2006:0411JUD004673399.

1 This structure follows the citation style of the ECtHR; see ECtHR (2022) Note explaining the mode of citation of the case-law of the Court and the Commission, available at <https://www.echr.coe.int/documents/d/echr/note_citation_eng> accessed 27 November 2024.

Erbakan v. Turkey, no. 59405/00, 6 July 2006, ECLI:CE:ECHR:2006:0706JUD005940500.
Erdoğdu and İnce v. Turkey [GC], nos. 25067/94 and 25068/94, ECHR 1999-IV, ECLI:CE:ECHR:1997:1211REP002506794.
Feldek v. Slovakia, no. 29032/95, ECHR 2001-VIII, ECLI:CE:ECHR:2001:0712JUD002903295.
Féret v. Belgium, no. 15615/07, 16 July 2009, ECLI:CE:ECHR:2009:0716JUD001561507.
Garaudy v. France (dec.), no. 65831/01, ECHR 2003-IX (extracts), ECLI:CE:ECHR:2003:0624DEC006583101.
Gündüz v. Turkey (dec.), no. 59745/00, ECHR 2003-XI (extracts), ECLI:CE:ECHR:2003:1204JUD003507197.
İ.A. v. Turkey, no. 42571/98, ECHR 2005-VIII, ECLI:CE:ECHR:2005:0913JUD004257198.
Identoba and Others v. Georgia, no. 73235/12, 12 May 2015, ECLI:CE:ECHR:2015:0512JUD007323512.
Ivanov v. Russia, no. 3436/05, 8 February 2007, ECLI:CE:ECHR:2007:0220DEC003522204.
Kaboğlu et Oran v. Turkey, nos. 1759/08, 50766/10 and 50782/10, 30 October 2018, ECLI:CE:ECHR:2018:1030JUD000175908.
Király and Dömötör v. Hungary, no. 10851/13, 17 January 2017, ECLI:CE:ECHR:2017:0117JUD001085113.
Le Pen v. France, no. 18788/09, 20 April 2010, ECLI:CE:ECHR:2010:0420DEC001878809.
Leroy v. France, no. 36109/03, 2 October 2008, ECLI:CE:ECHR:2008:1002JUD003610903.
Lilliendahl v. Iceland, no. 29297/18, 12 June 2018, ECLI:CE:ECHR:2020:0512DEC002929718.
M'Bala M'Bala v. France (dec.), no. 25239/13, ECHR 2015 (extracts), ECLI:CE:ECHR:2015:1020DEC002523913.
Matthews v. the United Kingdom [GC], no. 24833/94, ECHR 1999-I, ECLI:CE:ECHR:1999:0218JUD002483394.
Medya FM Reha Radyo ve ületiüim Hizmetleri A.ü. v. Turkey (dec.), no. 32842/02, 14 November 2006, ECLI:CE:ECHR:2006:1114DEC003284202.
Nix v. Germany, no. 38285/16, 13 Mar 2018, ECLI:CE:ECHR:2018:0313DEC003528516.
Norwood v. the United Kingdom (dec.), no. 23131/03, ECHR 2004-XI, ECLI:CE:ECHR:2004:1116DEC002313103.
Ottan v. France, no. 41841/12, 19 April 2018, ECLI:CE:ECHR:2018:0419JUD004184112.
Özgür Gündem v. Turkey, no. 23144/93, ECHR 2000-III, ECLI:CE:ECHR:2000:0316JUD002314493.
Pastörs v. Germany, no. 55225/14, 3 October 2019, ECLI:CE:ECHR:2019:1003JUD005522514.
Perinçek v. Switzerland [GC], no. 27510/08, ECHR 2015 (extracts), ECLI:CE:ECHR:2015:1015JUD002751008.
PETA Deutschland v. Germany, no. 43481/09, 8 November 2012, ECLI:CE:ECHR:2012:1108JUD004348109.
Roj TV A/S v. Denmark, no. 24683/14, 17 April 2018, ECLI:CE:ECHR:2018:0417DEC002468314.
Sanchez v. France [GC], no. 45581/15, 15 May 2023, ECLI:CE:ECHR:2023:0515JUD004558115.
Savva Terentyev v. Russia, no. 10692/09, 28 August 2018, ECLI:CE:ECHR:2018:0828JUD001069209.

Schimanek v. Austria, no. 32307/96, 1 February 2000, ECLI:CE:ECHR:2000:0201 DEC003230796.
Seurot v. France (dec), no. 57383/00, 18 May 2004, ECLI:CE:ECHR:2004:0518DEC 005738300.
Šimunić v. Croatia (dec), no. 20373/17, 22 January 2019, ECLI:CE:ECHR:2019:0122DEC 002037317.
Smajić v. Bosnia & Herzegovina, no. 48657/16, 16 January 2018, ECLI:CE:ECHR:2018: 0116DEC004865716.
Soulas and Others v. France, no. 15948/03, 10 July 2008, ECLI:CE:ECHR:2008:0710JUD 001594803.
Sousa Goucha v. Portugal, no. 70434/12, 22 March 2016, ECLI:CE:ECHR:2016:0322JUD 007043412.
Stomakhin v. Russia, no. 52273/07, 9 May 2018, ECLI:CE:ECHR:2018:0509JUD005227307.
Sürek and Özdemir v. Turkey [GC], nos. 23927/94 and 24277/94, 8 July 1999, ECLI:CE: ECHR:1999:0708JUD002392794.
Sürek v. Turkey (no. 1) [GC], no. 26682/95, ECHR 1999-IV, ECLI:CE:ECHR:1999:0708 JUD002668295.
Sürek v. Turkey (no. 4) [GC], no. 24762/94, 8 July 1999, ECLI:CE:ECHR:1999:0708JUD 002476294.
Vajnai v. Hungary, no. 33629/06, ECHR 2008, ECLI:CE:ECHR:2008:0708JUD003362906.
Vejdeland and Others v. Sweden, no. 1813/07, 9 February 2012, ECLI:CE:ECHR:2012:0209 JUD000181307.
Vereinigung Bildender Künstler v. Austria, no. 68354/01, 25 January 2007, ECLI:CE: ECHR:2007:0125JUD006835401.
W.P. and Others v. Poland (dec.), no. 42264/98, ECHR 2004-VII (extracts), ECLI:CE: ECHR:2004:0902DEC004226498.
Williamson v. Germany, no. 64496/17, 8 January 2019, ECLI:CE:ECHR:2019:0108DEC 006449617.

Judgements of the "old" Court (from 1960 until 31 October 1998)

· *Glimmerveen and Hagenbeek v.the Netherlands*, nos. 8348/78 and 8406/78, 11 October 1979, ECLI:CE:ECHR:1979:1011DEC000834878.
· *Pine Valley Developments Ltd and Others v. Ireland* (Article 50), 9 February 1993, Series A no. 246-B, ECLI:CE:ECHR:1993:0209JUD001274287.
· *Lawless v. Ireland*, no. 332/57, 1 July 1961, ECLI:CE:ECHR:1961:0701JUD000033257.
· *Handyside v. the United Kingdom*, 7 December 1976, Series A no. 24, ECLI:CE:ECHR: 1976:1207JUD000549372.
· *Tyrer v. the United Kingdom*, 25 April 1978, Series A no. 26, ECLI:CE:ECHR:1978: 0425JUD000585672.
· *Airey v. Ireland* (Article 50), 6 February 1981, Series A no. 41, ECLI:CE:ECHR:1981: 0206JUD000628973.
· *Case of Colossa v. Italy*, no. 9024/80, 12 February 1985, ECLI:CE:ECHR:1985: 0212JUD000902480.
· *Jersild v. Denmark*, 23 September 1994, Series A no. 298, ECLI:CE:ECHR:1994: 0923JUD001589089.

· *Honsik v. Austria*, no. 25062/94, 18 October 1995, ECLI:CE:ECHR:1995:1018DEC00250 6294.
· *Marais v. France*, no. 31159/96, 24 June 1996, ECLI:CE:ECHR:1996:0624DEC00311 5996.
· *Incal v. Turkey*, 9 June 1998, Reports of Judgments and Decisions 1998-IV, ECLI:CE: ECHR:1998:0609JUD002267893.
· *Nachtmann v. Austria*, no. 36773/97, 9 September 1998, ECLI:CE:ECHR:1998:0909 DEC003677397.
· *B.H., M.W., H.P. and G.K. v. Austria*, no. 12774/87, 12 October 1989.

Decisions and reports of the Commission

· *German Communist Party and Others v. Federal Republic of Germany*, no. 250/57, 20 July 1957.

Court of Justice of the European Union

· European Court of Justice, Opinion of Advocate General Szpunar delivered on 8 June 2023 (1) Case C-376/22, ECLI identifier: ECLI:EU:C:2023:467.
· Judgment of the Court of 12 November 1969, *Erich Stauder v. City of Ulm – Sozialamt*, Reference for a preliminary ruling: Verwaltungsgericht Stuttgart – Germany, Case 29-69, ECLI identifier: ECLI:EU:C:1969:57.
· Judgment of the Court of 17 December 1970. *Internationale Handelsgesellschaft mbH v Einfuhr- und Vorratsstelle für Getreide und Futtermittel*. Reference for a preliminary ruling: Verwaltungsgericht Frankfurt am Main – Germany. Case 11-70, ECLI identifier: ECLI:EU:C:1970:114.
· Judgment of the Court of 11 July 1985. Cinéthèque SA and others v Fédération nationale des cinémas français. References for a preliminary ruling: Tribunal de grande instance de Paris – France. Distribution of films in the form of video recordings – National prohibitions. Joined cases 60 and 61/84, ECLI identifier: ECLI: EU:C:1985:329.
· Judgment of the Court (Grand Chamber) of 21 December 2016, *Tele2 Sverige AB v. Post- och telestyrelsen and Secretary of State for the Home Department v. Tom Watson and Others*, Joined Cases C-203/15 and C-698/15, ECLI:EU:C:2016:970.
· Joined Cases C-682/18 and C-683/18: Judgment of the Court (Grand Chamber) of 22 June 2021 (requests for a preliminary ruling from the Bundesgerichtshof – Germany) – Frank Peterson v Google LLC, YouTube LLC, YouTube Inc., Google Germany GmbH (C-682/18) and Elsevier Inc. v Cyando AG (C-683/18), ECLI:EU:C: 2021:503.
· Judgment of the Court (Grand Chamber) of 4 July 2023, *Meta Platforms Inc and Others v. Bundeskartellamt*, ECLI:EU:C:2022:704, Opinion of Advocate General Rantos delivered on 20 September 2022 (1) case C-252/21, ECLI:EU:C:2023:537.

Other Relevant Case Law

· *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 29 August 2024.
· *Twitter, Inc. v. Taameh* (05/18/2023) United States Supreme Court, Twitter v. Taamneh 598 US (2023), available at <https://www.supremecourt.gov/opinions/22pdf/21-1496_d18f.pdf> accessed 12 January 2025.

Policy Documents

United Nations

· United Nations, High Commission for Human Rights, Annual Report 2023, A/HRC/22/17/Add.4 (Rabat Plan of Action).
· United Nations, High Commissioner for Human Rights, Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013).
· United Nations, High Commissioner for Human Rights: addendum, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013).
· United Nations, Human Rights Council, 'Report of the independent international fact-finding mission on Myanmar' (2018) A/HRC/39/64.
· United Nations, Human Rights Council, 'Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, John Ruggie' (2011) A/HRC/17/31 (UNGPs).
· United Nations, Human Rights Council, Report of the Special Rapporteur on minority issues, Report on hate speech, social media and minorities, (2021) (A/HRC/46/57).
· United Nations, Office of the High Commissioner for Human Rights (2011) Relevant Council of Europe Standards and Policies on the Prohibition and Prevention of "Hate Speech", Prepared by Directorate General of Human Rights and Legal Affairs (DGHL).
· United Nations, Office of the High Commissioner, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, A/RES/60/147.
· United Nations, Office of the High Commissioner, Implementing the UN "Protect, Respect and Remedy Framework" (UNGPs Guide).
· United Nations, Office of the High Commissioner, The Corporate Responsibility to Respect Human Rights, An Interpretative Guide.
· United Nations, Report of the Special Rapporteur on minority issues (2021) A/HRC/46/57.
· United Nations, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/74/486 (May 7, 2010).
· United Nations, Resolution adopted by the General Assembly on 16 December 2005, Basic Principles and Guidelines on the Right to a Remedy and Reparation for Victims of Gross Violations of International Human Rights Law and Serious Violations of International Humanitarian Law, A/RES/60/147, II(3).

· United Nations, Secretary-General, Globalization and Its Impact on the Full Enjoyment of All Human Rights", U.N. Doc. A/55/342 (Aug. 31, 2000).
· United Nations, 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 (Aug 26, 2003).
· United Nations, World Conference on Human Rights, Vienna Declaration and Programme of Action, 1 30, U.N. Doe. A/CONF. 157/23 (June 25, 1993).

Organisation for Economic Co-operation and Development (OECD)

· OECD, 'OECD Guidelines for Multinational Enterprises' (2011) available at <http://mneguidelines.oecd.org/guidelines/> accessed 6 April 2023.
· 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.
· 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.
· OECD, 'Business and Finance Outlook 2021' available at <https://www.oecd.org/finance/oecd-business-and-finance-outlook-26172577.htm> accessed 6 April 2023.

COUNCIL OF EUROPE

Committee of Ministers, Recommendations

· Council of Europe, Committee of Ministers, Recommendation No. R (97) 20 of the Committee of Ministers to member states on "hate speech", Adopted on 30 October 1997.
· Council of Europe, Committee of Ministers, Recommendation No. R (97) 21 of the Committee of Ministers to member states on the media and the promotion of a culture of tolerance, Adopted on 30 October 1997.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2010)5 of the Committee of Ministers to member states on measures to combat discrimination on grounds of sexual orientation or gender identity, Adopted on 31 March 2010.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2011)7 of the Committee of Ministers to member states on a new notion of media, Adopted on 21 September 2011.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a Guide to human rights for Internet users, Adopted on 16 April 2014.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2016)3 of the Committee of Ministers to member States on human rights and business, Adopted on 2 March 2016.

· Council of Europe, Committee of Ministers Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the roles and responsibilities of internet intermediaries, Adopted on 7 March 2018.
· Council of Europe, Committee of Ministers, 'Declaration by the Committee of Ministers on the manipulative capabilities of algorithmic processes' (2019), Adopted on 13 February 2019.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems, Adopted on 8 April 2020.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2022)16https: //search.coe.int/cm – _ftn1 of the Committee of Ministers to member States on combating hate speech, Adopted on 20 May 2022.
· Council of Europe, Committee of Ministers, Recommendation CM/Rec(2024)4 of the Committee of Ministers to member States on combating hate crime, Adopted on 7 May 2024.

Council of Europe, Recommendations' Explanatory Memoranda and Other Documents

· Council of Europe, Ministers' Deputies CM Documents CM(2014)31-addfinal, Recommendation CM/Rec(2014)6 of the Committee of Ministers to member States on a guide to human rights for Internet users – Explanatory Memorandum, Adopted on 16 April 2014.
· Council of Europe, Ministers' Deputies, CM Documents, CM(2005)80 final 17 May 2005, Action Plan.
· Council of Europe, European Ministerial Conferences on Mass Media Policy and Council of Europe Conferences of Ministers responsible for Media and New Communications Services, Strasbourg 2021, available at <https://rm.coe.int/168064 61f b> accessed 29 August 2024.
· Council of Europe, Committee of Ministers, Guidelines of the Committee of Ministers of the Council of Europe on upholding equality and protecting against discrimination and hate during the Covid-19 pandemic and similar crises in the future (2021), Adopted on 5 May 2021.
· Council of Europe, Steering Committee on Anti-Discrimination, Diversity and Inclusion (CDADI) and Steering Committee on Media and Information Society (CDMSI) – Recommendation CM/Rec(2022)16 of the Committee of Ministers to member States on combating hate speech – Explanatory Memorandum, Adopted 20 May 2022.

European Commission against Racism and Intolerance (ECRI)

· Council of Europe, European Commission against Racism and Intolerance (ECRI), General Policy Recommendation (GPR) No. 6 on combating the dissemination of racist, xenophobic and antisemitic material via the Internet, adopted on 15 December 2000.
· Council of Europe, European Commission against Racism and Intolerance (ECRI), General Policy Recommendation (GPR) No. 7 on national legislation to combat

racism and racial discrimination, Adopted on 13 December 2002 and Amended on 7 December 2017.
· Council of Europe, European Commission against Racism and Intolerance (ECRI), General Policy Recommendation (GPR) No. 11 on combating racism and racial discrimination in policing, Adopted on 29 June 2007.
· Council of Europe, European Commission against Racism and Intolerance (ECRI), General Policy Recommendation (GPR) No. 15 on combating hate speech, Adopted on 8 December 2015.

Other Council of Europe Policy Documents

· Council of Europe, Keynote speech of Nils Muiz?nieks, Council of Europe Commissioner for Human Rights, 'Freedom of Expression and Democracy in the Digital Age: Opportunities, Rights, Responsibilities' (November 7-8, 2013) available at <https://www.statewatch.org/media/documents/news/2013/nov/coe-speech-freedom-of-expression.pdf> accessed 6 April 2023.

EUROPEAN UNION

European Commission

· European Union, European Commission, Directorate-General (DG) Justice, Guidance Document related to the transposition and implementation of Directive 2012/29/EU of the European Parliament and of the Council of 25 October 2012 establishing minimum standards on the rights, support and protection of victims of crime, and replacing Council Framework Decision 2001/220/JHA, 12.2013, available at <https://commission.europa.eu/document/download/238cafb6-d5cd-4d1a-8624-a0bafb2cdfa3_en?filename=13_12_19_3763804_guidance_victims_rights_directive_eu_en.pdf> accessed 29 August 2024.
· European Union, European Commission, Report from the Commission to the European Parliament and the Council on the implementation of Council Framework Decision 2008/913/JHA on combating certain forms and expressions of racism and xenophobia by means of criminal law, /* COM/2014/027 final */, 27.1.2014.
· European Union, European Commission, Communication from the Commission Guidelines on the practical application of the essential functionality criterion of the definition of a 'video-sharing platforms service' under the Audiovisual Media Services Directive 2020/C 223/02, OJ C 223, 7.7.2020, p. 3–9.
· European Union, Communication from the Commission to the European Parliament and the Council, A more inclusive and protective Europe: extending the list of EU crimes to hate speech and hate crime, COM/2021/777 final, Brussels, 9.12.2021.

European Parliament

· European Union, European Parliament resolution of 18 January 2024 on extending the list of EU crimes to hate speech and hate crime (2023/2068(INI)), available at

<https://www.europarl.europa.eu/doceo/document/TA-9-2024-0044_EN.pdf> accessed 13 of January 2025.

Policy Supporting Documents

United Nations

- United Nations, Definition of Gross and Large-scale Violations of Human Rights as an International Crime, Comm. on Human Rights, Prevention of Discrimination and Protection of Minorities, Working paper submitted by Mr. Stanislav Chemichenko in accordance with Sub-Commission decision 1992/109, 14, U.N. Doc. E/CN.4/Sub.2/1993/10, 08.06.1993.
- United Nations, General Assembly, Independent International Fact-Finding Mission on Myanmar, 'Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar' (IIMM, Detailed findings), A/HRC/39/CRP.2, 12. 09.2018.
- United Nations, United Nations Women, Frequently Asked Questions: Trolling, stalking, doxing and other forms of violence against women in the digital age, 19.11.2024, available at <https://www.unwomen.org/en/what-we-do/ending-violence-against-women/faqs/tech-facilitated-gender-based-violence> accessed 18 November 2024.

Council of Europe

- Council of Europe, European Court of Human Rights, Press Release on Admissibility Decision of Le Pen v. France (App. No. 18788/2009), 07.05.2010.
- Council of Europe, Explanatory Report, Convention on Preventing and Combating Violence Against Women and Domestic Violence, May 11, 2011, E.T.S. 210, Istanbul, 11.V.2011.
- Council of Europe, European Court of Human Rights, Note explaining the mode of citation of the case-law of the Court and the Commission, Updated October 2022, available at <https://www.echr.coe.int/documents/d/echr/note_citation_eng> accessed 27 November 2024.
- Council of Europe, European Court of Human Rights, Guide on Article 7 of the European Convention on Human Rights, No punishment without law: the principle that only the law can define a crime and prescribe a penalty, updated on 2022, available at <https://ks.echr.coe.int/documents/d/echr-ks/guide_art_7_eng> accessed 11 April 2024.
- Council of Europe, 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 11 April 2024.
- Council of Europe, European Court of Human Rights, Factsheet – Hate Speech, Press Unit, November 2023.
- Council of Europe, European Court of Human Rights, Guide on Article 13 of the ECHR Right to an effective remedy, Updated on 31 August 2024, available at

<https://ks.echr.coe.int/documents/d/echr-ks/guide_art_13_eng> accessed 28 May 2024.

European Union

European Commission

- European Union, European Commission, Didier Reynders, Directorate-General for Justice and Consumers, 5th valuation of the Code of Conduct on Countering Illegal Hate Speech Online, June 4, 2020.
- European Union, European Commission, Press Release, EU Code of Conduct against illegal hate speech online: results remain positive but progress slows down, October 17, 2021, available at <https://ec.europa.eu/commission/presscorner/detail/en/ip_21_5082> accessed 29 August 2024.
- European Union, European Commission, Press Release, IP/22/1533, International Women's Day 2022: Commission Proposes EU-Wide Rules to Combat Violence Against Women and Domestic Violence, March 8, 2022.
- European Union, European Commission, Press Release, 'Digital Services Act: Commission designates first set of VLOPs and Search Engines', April 25, 2023, available at <https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413> accessed 29 August 2024.
- European Union, European Commission, Press Release, 'No place for hate in Europe. Commission and High Representative launch call to action to unite against all forms of hatred', December 6, 2023, available at <https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6329> accessed 29 August 2024.
- European Union, European Commission, Questions and answers on the Digital Services Act, February 23, 2024, available at <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348> accessed 29 August 2024.
- European Union, European Commission, Communications Networks, Content and Technology, Connect develops and implements policies to make Europe fit for the digital age, available at <https://commission.europa.eu/about/departments-and-executive-agencies/communications-networks-content-and-technology_en> accessed 26 August 2024.
- European Union, European Commission, Monitoring rounds of the Code of conduct on countering illegal hate speech online, available at <https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 10 April 2024.

European Parliament

- European Union, European Parliament, Legislative Train Schedule, Proposals to extend the list of EU crimes to all forms of hate crime and hate speech, 15.12.2024, available at <https://www.europarl.europa.eu/legislative-train/theme-a-new-push-for-european-democracy/file-hate-crimes-and-hate-speech> accessed 13 January 2025.

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

Other EU Policy Supporting Documents

· European Union, EU Network of Independent Experts on Fundamental Rights, Commentary of the Charter of Fundamental Rights of the European Union, June 2006, available at <https://sites.uclouvain.be/cridho/documents/Download.Rep/NetworkCommentaryFinal.pdf> accessed 13 January 2025.
· European Union, EUROJUST, Third report of the observatory function on encryption, July 2, 2021, available at <https://www.eurojust.europa.eu/publication/third-report-observatory-function-encryption> accessed 21 February 2024.
· 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.
· European Union, Agency for Fundamental Rights, Press Release, Harassment and violence against LGBTIQ people on the rise, May 14, 2024, available at <https://fra.europa.eu/en/news/2024/harassment-and-violence-against-lgbtiq-people-rise> accessed 18 November 2024.

Academic Literature

Alegre, Susie (2022) *Freedom to Think: The Long Struggle to Liberate Our Minds*. Atlantic Books.

Alkiviadou, Natalie (2018) 'The Legal Regulation of Hate Speech: The International and European Frameworks,' *55 POLITIÈKA MISAO* 203, 223.

Alkiviadou, Natalie (2019) 'Hate speech on social media networks: towards a regulatory framework?.' *Information & Communications Technology Law* 28.1: 19-35.

Appelman, Naomi, João Pedro Quintais and Ronan Fahy (2022), 'Using Terms and Conditions to apply Fundamental Rights to Content Moderation', *German Law Journal 24.5 (2023): 881-911*.

Argyrou, Aikaterini (2017) 'Making the case for case studies in empirical legal research.' *Utrecht Law Review* 13.3: 95-113.

Bartlett, Katharine T (2018) 'Feminist Legal Methods,' *103 Harvard Law Review* 829-888.

Bartlett, Katharine T and Rosanne Kennedy (1991) *Feminist Legal Theory, Readings in Law and Gender*, Routledge, 460.

Bayer, Judit & Petra Bard (2020), 'Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches,' *Study requested by the LIBE committee, Policy Department for Citizens' Rights and Constitutional Affairs Directorate-General for Internal Policies PE 655.135* – July 2020, available at <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/655135/IPOL_STU(2020)655135_EN.pdf > accessed 21 Feb 2024.

Bayer, Judit, Bernd Holznagel, Päivi Korpisaari (ex. Tiilikka), Lorna Woods, (2021) *Perspectives on Platform Regulation: Concepts and Models of Social Media Governance*

Across the Globe, Volume 1, Baden-Baden, Nomos, 601, available at <https://doi.org/10.5771/9783748929789> accessed 29 August 2024.

Bell, Derrick (2008), *And We Are Not Saved: The Elusive Quest For Racial Justice*, Basic Books.

Ben-David, Anat and Ariadna Matamoros Fernández (2016), 'Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain'. *International Journal of Communication*, 10, 27.

Bertolini, Andrea et al. (2021), 'Liability of Online Platform: Study for the European Parliament' *European Parliamentary Research Service* PE 656.318.

Bodig, Matyas (2015), 'Legal Doctrinal Scholarship and Interdisciplinary Engagement,' *Erasmus Law Review,* Vol. 8: 43.

Boyle, Kevin & Anneliese Baldaccini (2017) *A Critical Evaluation of International Human Rights Approaches to Racism*, 1st Edition, 2011, Routledge, 57.

Bradford, Anu (2020), *The Brussels Effect: How the European Union Rules the World*, Oxford University Press, 27 February 2020.

Braman, Sandra and Stephanie Roberts (2003) 'Advantage ISP: Terms of service as media law.' *New media & society* 5.3: 422-448.

Buri, Ilaria and Joris van Hoboken (2021) 'The Digital Services Act (DSA) proposal: a critical overview.' *Digital Services Act (DSA) Observatory, Discussion paper* – version of 28 October 2021, available at <https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf> accessed 13 January 2025.

Bursztein, Elle, Clarke, E., DeLaune, M., Elifff, D. M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K., & Bright, T. (2019). 'Rethinking the Detection of Child Sexual Abuse Imagery on the Internet.' *The World Wide Web Conference*, 2601–2607, available at <https://doi.org/10.1145/3308558.3313482> accessed 7 Sep 2023.

Carastathis, Anna (2016), *Intersectionality: Origins, Contestations, Horizons*, University of Nebraska Press, 312.

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

Carlson, Caitlin Ring (2021), *Hate Speech*, MIT Press Essential Knowledge series, April 6, 2021.

Chitra, Uthsav and Christopher Musco (2020) 'Analyzing the impact of filter bubbles on social network polarization.' *Proceedings of the 13th International Conference on Web Search and Data Mining*.

Cohen-Almagor, Raphael (2019), Racism and Hate Speech – A Critique of Scanlon's Contractual Theory, 53 First Amendment Studies 1, 2.

Combs, Gene (2019) 'White privilege: what's a family therapist to do?' *Journal of marital and family therapy* 45.1: 61-75.

Cooper, Anna Julia (1892), *A Voice From The South*, Xenia, Ohio, The Aldine Printing House, 1892.

Crawford, Kate (2021) *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*, Yale University Press.

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

Crenshaw, Kimberlé (2013) Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics, *Feminist legal theories,* Routledge, 2013, 23-51.

De Hert, Paul and Serge Gutwirth (2006) 'Privacy, data protection and law enforcement. Opacity of the individual and transparency of power.' *Privacy and the criminal law:* 61-104.

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

Delgado, Richard and Jean Stefancic (2004) *Understanding Words That Wound*, Routledge 12–19, available at <https://doi.org/10.4324/9780429503351> accessed 29 August 2024.

Delmonaco, Daniel et al. (2024) '"What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and" Prove" Shadowbanning."' *Proceedings of the ACM on Human-Computer Interaction 8*.CSCW1: 1-39.

Deva, Surya (2012) 'Guiding Principles on Business and Human Rights: Implications for Companies' 9(2) *European Company Law* 101.

Deva, Surya (2023) 'Mandatory human rights due diligence laws in Europe: A mirage for rightsholders?' *Leiden Journal of International Law.* 2023;36(2):389-414.

Di Gangi, Paul M and Molly M. Wasko (2016) 'Social media engagement theory: Exploring the influence of user engagement on social media usage.' *Journal of Organizational and End User Computing (JOEUC)* 28.2: 53-73.

Dias Oliva, Thiago, Dennys Marcelo Antonialli and Alessandra Gomes (2021) 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online', *Sexuality & Culture* 25, 700.

Dittel, Alexander (2022) 'The UK's Online Safety Bill: The day we took a stand against serious online harms or the day we lost our freedoms to platforms and the state?.' *Journal of Data Protection & Privacy* 5.2: 183-194.

Dragiewicz, Molly, Jean Burgess, Ariadna Matamoros-Fernández, Michael Salter, Nicolas P. Suzor, Delanie Woodlock, and Bridget Harris (2018) 'Technology facilitated coercive control: Domestic violence and the competing roles of digital media platforms' *Feminist Media Studies*, 18(4), 609-625.

Dworkin, Ronald Myles (1999) *Freedom's law: the moral reading of the American constitution*, Oxford University Press, 1999.

El-Tayeb, Fatima (2011) *European Others, Queering Ethnicity in Postnational Europe*, University of Minnesota Press, 2011.

Feagin, Joe R. and Debra Van Ausdale (2001) *The first R: How children learn race and racism*, Rowman & Littlefield Publishers.

Fourie, Andria Naudé (2015) "Expounding the place of legal doctrinal methods in legal-interdisciplinary research." *Erasmus Law Review* 8: 95.

Françoise Tulkens (2013) 'The hate factor in political speech: Where do responsibilities lie?', *Report of the Council of Europe Conference on "The hate factor in political speech: Where do responsibilities lie?"*, Warsaw, 18-19 September 2013, Doc. No. MCM(2013)002.

Galvan, Justice Belen (2020) "Facebook's Legal Responsibility for the Rohingya Genocide". *USFL Rev.*, 55, 123.

Gelber, Katharine (2002). *Speaking Back. The free speech versus hate speech debate.* John Benjamins Publishing Company, 117.

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

Gelber, Katharine (2021) 'Differentiating hate speech: a systemic discrimination approach.' *Critical Review of International Social and Political Philosophy*, 24(4), 393–414.

Genç-Gelgeç, Berrak (2022) 'Regulating Digital Platforms: Will the DSA Correct Its Predecessor's Deficiencies?' 18 *Croatian Yearbook of European Law and Policy* 25.

Gerads, Janneke (2013) 'How to improve the necessity test of the European Court of Human Rights.' *International Journal of Constitutional Law* 11.2: 466-490.

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

Gillespie, Tarleton (2020) 'Content moderation, AI and the question of scale,' *Big Data and Society*, 7.2 (2020): 2053951720943234.

Gjøsteen, Kristian, Thomas Haines, Johannes Müller, Peter Rønne, and Tjerand Silde (2022) 'Verifiable decryption in the head', *Australasian Conference on Information Security and Privacy*, Springer International Publishing, 355-374.

Goldman, Eric (2021) 'Content moderation remedies' *Mich. Tech. L. Rev*. 28: 1., 24.

Gonzaìlez-Fuster, Gloria, Rosamunde Van Brakel, and Paul De Hert (Eds.) (2022), *Research handbook on privacy and data protection law: values, norms and global politics*, Edward Elgar Publishing.

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

Greschbach, Benjamin, Kreitz, G., & Buchegger, S. (2012). 'The devil is in the metadata – New privacy challenges in Decentralised Online Social Networks.' *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, 333–339, available at <https://doi.org/10.1109/PerComW.2012.6197506> accessed 7 Sep 2023.

Griffin, Rachel (2023) 'Rethinking rights in social media governance: human rights, ideology and inequality' *European Law Open* 2.1: 30-56.

Griffin, Rachel (2023). 'The Law and Political Economy of Online Visibility: Market Justice in the Digital Services Act.' *Technology & Regulation*, 2023, 69-79. available at <https://doi.org/10.26116/techreg.2023.007> accessed 28 August 2024.

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

Haimson, Oliver L. et al. (2021) 'Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas.' *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2: 1-35.

Hakim, Neema (2020) 'How social media companies could be complicit in incitement to genocide.' *Chi. J. Int'l L.* 21: 83.

Hare, Ivan & James Weinstein (2009), *Extreme Speech And Democracy*, Oxford University Press, 720.

Heller, Brittan and Joris van Hoboken (2019) 'Freedom of Expression: A Comparative Summary of United States and European Law' *Transatlantic High Level Working*

Group on Content Moderation Online and Freedom of Expression Working Paper, SSRN 4563882 (2019).

Hoecke, Mark Van (2015) *Methodology of Comparative Legal Research, Law and Method*, *Pravovedenie* (2013): 121.

Hörnle, Julia (2021). *Internet jurisdiction law and practice*. Oxford University Press.

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

Husovec, Martin and Irene Roche Laguna (2023) *Digital services act: A short primer*, Martin Husovec and Irene Roche Laguna, Principles of the Digital Services Act, Oxford University Press.

Hutchinson, M. R. (1999) 'The Margin of Appreciation Doctrine in the European Court of Human Rights,' *The International and comparative law quarterly*, 48 (3), 638–650.

Jacobs, James B. & Kimberly Potter (2001), *Hate Crimes: Criminal Law And Identity Politics*, Oxford University Press, 2000.

Jägers, Nicola (2011) 'UN Guiding Principles on Business and Human Rights: Making headway towards real corporate accountability?' 29(2) *Netherlands Quarterly of Human Rights* 159–163.

Jeffrey, James (2020). 'The smart feature phone revolution in developing countries: Bringing the internet to the bottom of the pyramid.' *The Information Society*, 36(4), 226–235.

Joseph, Sarah and Joanna Kyriakakis (2023) 'From soft law to hard law in business and human rights and the challenge of corporate power' 36(2) *Leiden Journal of International Law* 335.

Kaye, David (2019) *Speech police: The global struggle to govern the Internet*, Columbia Global Reports, 144.

Khosravi, Ooryad, S. (2023). 'Alt-right and authoritarian memetic alliances: global mediations of hate within the rising Farsi manosphere on Iranian social media.' *Media, Culture and Society* 45.3 (2023): 487-510.

Klamberg, Mark, ed. (2017) *Commentary on the law of the International Criminal Court*. Vol. 29. Torkel Opsahl Academic EPublisher.

Klonick, Kate (2017) 'The new governors: The people, rules, and processes governing online speech.' *Harv. L. Rev.* 131: 1598.

Klonick, Kate (2019) 'The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression.' *Yale LJ* 129: 2418.

Koehler, Daniel (2022) 'From superiority to supremacy: Exploring the vulnerability of military and police special forces to extreme right radicalization' *Studies in Conflict & Terrorism*: 1-24.

Kosinova, Darina S. and Arsenii V. Paliiuk (2021) 'Prohibition of Discrimination: Concepts, Features and Obligations of the State according to the Convention for the Protection of Human Rights and Fundamental Freedoms'. *L. & Innovative Soc'y*, 99.

Labey, Eline and Valentina Golunova (2022). 'Judges of Online Legality: Towards Effective User Redress in the Digital Environment' *In European Yearbook on Human Rights* (1 ed., pp. 105-135). Intersentia.

Laidlaw, Emily B. (2012) The responsibilities of free speech regulators: an analysis of the Internet Watch Foundation, *International Journal of Law and Information Technology*.

Lane, Lottie (2021) 'A Human Rights Responsibility Primer for Businesses Developing AI: Part 2', *Medium*, 14 September 2021.

Lane, Lottie (2022) 'Clarifying Human Rights Standards through Artificial Intelligence Initiatives: A multi-level comparative analysis' *International and Comparative Law Quarterly* 74(1) 16.

Lane, Lottie (2023) 'Artificial Intelligence and Human Rights: Corporate responsibility in AI governance initiatives' *Nordic Journal of Human Rights*.

Lane, Lottie (2023) '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*. Cham: Springer International Publishing, 2023. 183-205.

Lane, Lottie (2023), 'Preventing long-term risks to human rights in smart cities: a critical review of responsibilities for private developers of AI', *Internet Policy Review* 12.1 (2023).

Lawrence, Charles R. (1987), 'The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism,' 39 *STAN. L. REV*. 317, 201.

Lawrence, Frederick M. (1994) 'The Punishment of Hate: Toward a Normative Theory of Bias-Motivated Crimes,' 93 *MICH. L. REV*. 320, 342–343.

Lee, Diana (2017) 'Germany's NetzDG and the threat to online free speech.' *Yale Media Freedom & Information Access Clinic Case Disclosed Blog*, October 10, 2017, available at <https://law.yale.edu/mfia/case-disclosed/germanys-netzdg-and-threat-online-free-speech> accessed 13 January 2025.

Leerssen, Paddy (2020) 'The soap box as a black box: Regulating transparency in social media recommender systems.' *European Journal of Law and Technology* 11.2.

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

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

Llorens, A., Tzovara, A., Bellier, L., Bhaya-Grossman, I., Bidet-Caulet, A., Chang, W. K., ... & Dronkers, N. F. (2021). 'Gender bias in academia: A lifetime problem that needs solutions,' *Neuron*, *109*(13), 2047-2074.

Locke, John, James H. Tully, and James Tully. *A letter concerning toleration*. Indianapolis: Hackett Publishing Company, 1983.

Love, Heather, 'Queer.' *Transgender studies quarterly* Volume1, Numbers 1-2 (2014): 172-176.

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

Ludovic, Terren and Rosa Borge-Bravo Rosa Borge-Bravo (2021) 'Echo chambers on social media: A systematic review of the literature.' *Review of Communication Research* 9.

Macchi, Chiara and Claire Bright (2020) '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 218.

Machado, Caio CV and Thaís Helena Aguia, (2023) 'Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models.' *Business and Human Rights Journal* 8.2: 244-251.

Machado, CCV, Aguiar TH (2023), Emerging Regulations on Content Moderation and Misinformation Policies of Online Media Platforms: Accommodating the Duty of Care into Intermediary Liability Models. *Business and Human Rights Journal*, 8(2):244-251.

Mackinnon, Catharine A. (1993), *Only Words,* Harvard University Press, 1993.

Matamoros-Fernández, Ariadna (2017) 'Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube.' *Information, Communication & Society* 20.6: 930-946.

Matamoros-Fernández, Ariadna and Johan Farkas (2021) 'Racism, hate speech, and social media: A systematic review and critique.' *Television & new media* 22.2: 205-224.

Mathew, Binny, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee (2019) 'Spread of Hate Speech in Online Media,' *Proceedings 10th ACM Conf. on Web Science*.

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

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

McConville, Mike (2007) (ed) *Research Methods for Law*, Edinburgh University Press, 2017

McCorquodale, Robert and Justine Nolan (2021) 'The Effectiveness of Human Rights Due Diligence for Preventing Business Human Rights Abuses' 68 *Netherlands International Law Review* 455.

McCoy, Henrika (2020), 'Black Lives Matter, and Yes, You Are Racist: The Parallelism of the Twentieth and Twenty-First Centuries,' 37 *CHILD ADOLESC. SOC. WORK J.* 463, 464.

McGonagle, Tarlach (2011) 'Minority Rights, Freedom of Expression and of the Media: Dynamics and Dilemmas' *Human Rights Research Series*, Vol. 44. Cambridge: Intersentia, 2011.

McGonagle, Tarlach (2019) '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*', Cambridge, Massachusetts and London, England, The MIT Press, 2019, pp. 227-253.

McGonagle, Tarlach (2020) 'Free expression and respect for others', in Yasha Lange, Ed., Living Together: *a handbook on Council of Europe standards on media's contribution to social cohesion, intercultural dialogue, understanding, tolerance and democratic participation* (Strasbourg, Council of Europe Publishing, 2009), pp. 7-20.

McGonagle, Tarlach, 'The Council of Europe against online hate speech: Conundrums and challenges', Expert paper, doc.no. MCM 2013(005), *the Council of Europe Conference of Ministers responsible for Media and Information Society, 'Freedom of*

Expression and Democracy in the Digital Age: Opportunities, Rights, Responsibilities', Belgrade, 7-8 November 2013.

Mill, John Stuart (1859*) On Liberty And Other Essays*, John Gray ed., *Oxford Univ. Press* 1998.

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

Moore, Martin and Tambini Damian (eds), (2022) *Regulating Big Tech: Policy Responses to Digital Dominance*, Oxford University Press, 2022.

Müller, Karsten & Carlo Schwarz (2021) 'Fanning the Flames of Hate: Social Media and Hate Crime,' 19 *Journal of the European Economic Association*, 2131.

Murray, Andrew D. (2011) 'Nodes and gravity in Virtual Space,' 208, *Legisprudence* 5.2 (2011): 195-221.

Myers West, S. (2018). 'Censored, suspended, shadow banned: User interpretations of content moderation on social media platforms.' *New Media & Society*, 20(11), 4366-4383.

O'Brien, Claire Methven and Jaques Hartmann (2022) 'The European Commission's proposal for a Directive on corporate sustainability due diligence: two paradoxes', *EJIL: Talk! Blog of the European Journal of International Law*, 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.

Parekh, Bhikhu (2012) 'Is There a Case for Banning Hate Speech?,' 37 – 56, in Michael Herz & Peter Molnar (eds.). *The Content And Context Of Hate Speech: Rethinking Regulation And Responses*, Cambridge University Press, 2012.

Pasquale, Frank. (2015) *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

Pentney, Katie (2021) 'Licensed to kill… discourse? Agents provocateurs and a purposive right to freedom of expression' *Netherlands Quarterly of Human Rights*, Vol. 39(3) 241-27.

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

Pollicino, Oreste & Gabriella Romeo (Eds.), (2016), *The internet and constitutional law?: the protection of fundamental rights and constitutional adjudication in Europe*. Routledge, Taylor & Francis Group.

Pollicino, Oreste (2021) *Judicial Protection of Fundamental Rights on the Internet: A Road Towards Digital Constitutionalism?* (1st ed.). Hart Publishing.

Quattrociocchi, Walter, Antonio Scala, and Cass R. Sunstein (2016) 'Echo chambers on Facebook.' Available at *SSRN* <https://ssrn.com/abstract=2795110> or <http://dx.doi.org/10.2139/ssrn.2795110> accessed 13 January 2025.

Quintais, João Pedro, Naomi Appelman, and Ronan Ó. Fathaigh, (2023) 'Using terms and conditions to apply fundamental rights to content moderation' *German Law Journal* 24.5: 881-911.

Rahalkar, Chaitanya, and Anushka Virgaonkar. "SoK: Content Moderation Schemes in End-to-End Encrypted Systems." *arXiv preprint arXiv:2208.11147* (2022).

Rapp, Kyle (2021) 'Social media and genocide: The case for home state responsibility.' *Journal of Human Rights* 20.4: 486-502.

Rawls, Anne Warfield & Waverly Duck (2020) *Tacit Racism*, The University of Chicago Press, 248.

Robert C Clark (1981) 'The interdisciplinary study of legal evolution.' *The Yale Law Journal* 90.5: 1238-1274.

Rosenfeld, Michel (2002) 'Hate Speech in Constitutional Jurisprudence: A Comparative Analysis', 24 *CARDOZO L. REV*. 1523, 1565.

Rubin, Edward L (1997) 'Law and the Methodology of Law.' *Wis. L. Rev*.: 521.

Scanlon, Thomas (1979) 'Freedom of Expression and Categories of Expression Principles of Expression and Restriction: A First Amendment Symposium,' 40 *U. PITT. L. REV*. 519, 527.

Scanlon, Thomas (2018) 'A Framework for Thinking about Freedom of Speech and some of its Implications,' UC Berkeley Law, available at <https://www.law.berkeley.edu/wp-content/uploads/2018/10/Freedom-of-Speech-Berkeley.pdf> accessed 13 January 2025.

Schauer, Frederick (1981) 'Categories and the First Amendment: A Play in Three Acts,' 34 *VAND. L. REV.* 265, 270.

Scheffler, Sarah, and Jonathan Mayer (2023) 'Sok: Content moderation for end-to-end encryption.' *arXiv preprint arXiv:2303.03979*.

Scheffler, Sarah, Anunay Kulshrestha, and Jonathan Mayer. 'Public verification for private hash matching.' *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023.

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

Spohr, Dominic (2017) 'Fake news and ideological polarization: Filter bubbles and selective exposure on social media.' *Business information review* 34.3: 150-160.

Stefancic, Jean, and Richard Delgado, eds. (2000) *Critical race theory: The cutting edge*. Philadelphia, PA: Temple University Press, 2000.

Stoica, Victor (2021) *Remedies before the International Court of Justice*. Cambridge University Press, 2021.

Suzor, Nicolas (2020) 'Understanding content moderation systems: new methods to understand internet governance at scale, over time, and across platforms.' In *Computational Legal Studies*, pp. 166-189. Edward Elgar Publishing.

Suzor, Nicolas P., Sarah Myers West, Andrew Quodling, and Jillian York. (2019) 'What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation.' *International Journal of Communication* 13: 18.

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

Terren, Ludovic Terren Ludovic, and Rosa Borge-Bravo Rosa Borge-Bravo (2021) 'Echo chambers on social media: A systematic review of the literature.' *Review of Communication Research* 9.

Trengove, Markus et al. (2022) 'A critical review of the Online Safety Bill.' *Patterns* 3.8.

Tulkens, Françoise (2012) 'When to say is to do: Freedom of expression and hate speech in the case-law of the European Court of Human Rights', European Court of Human Rights – European Judicial Training Network, *Seminar on Human Rights for European Judicial Trainers*, 7 July 2015.

Turillazzi, Aina, Mariarosaria Taddeo, Luciano Floridi, and Federico Casolari. (2023) 'The digital services act: an analysis of its ethical, legal, and social implications.' *Law, Innovation and Technology* 15.1: 83-106.

Tworek, Heidi, and Paddy Leerssen. 'An analysis of Germany's NetzDG law.' *First session of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression* (2019).

Udupa, Sahana. 'Decoloniality and extreme speech.' *65th e-seminar, Media Anthropology Network, European Association of Social Anthropologists*. 2020.

Viejo Otero, Paloma. *Governing hate: Facebook and hate speech*. Diss. Dublin City University, 2022.

Waldron, Jeremy (2014) *The Harm In Hate Speech*, Harvard University Press.

Webley, Lisa (2010) 'Qualitative approaches to empirical legal research', In Peter Cane & Herbert M. Kritzer (eds.), *The Oxford handbook of empirical legal research*. New York: Oxford University Press (2010).

William E. Donald. (2024) 'Merit beyond metrics: Redefining the value of higher education. Industry and Higher Education; Robin Cowan, Moritz Müller, Alan Kirman, Helena Barnard, Overcoming a legacy of racial discrimination: competing policy goals in South African academia,' *Socio-Economic Review*, Volume 22, Issue 3, July 2024, Pages 1413–1449.

Williams, M. T. (2019) 'Adverse racial climates in academia: Conceptualization, interventions, and call to action.' *New Ideas in Psychology*. 55:58–67.

Witting, Sabine (2020). *Child sexual abuse in the digital era?: Rethinking legal frameworks and transnational law enforcement collaboration*, available at <https://scholarlypublications.universiteitleiden.nl/access/item%3A2966707/view> accessed 13 January 2025.

Witting, Sabine and Mark Leiser (2023) 'Outcome Report of the 2nd Expert Workshop on EU proposed Regulation on Preventing and Combatting Child Sexual Abuse,' Leiden University, available at <https://www.universiteitleiden.nl/binaries/content/assets/rechtsgeleerdheid/instituut-voor-metajuridica/final-eu-workshop-report-csa-proposal-2nd-workshop-05042023.pdf> accessed 17 October 2023.

Witting, Sabine and Mark Leiser (2023) 'Outcome Reports of the 1st expert Workshop on Eu proposed Regulation on Preventing and Combatting Child Sexual Abuse Council of Europe, available at <https://rm.coe.int/outcome-report-of-the-expert-workshop-on-eu-proposed-regulation-on-pre/1680aa00e4> accessed 17 October 2022.

Witting, Sabine K. and Gianclaudio Malgieri (2023) 'Voluntary detection order under the proposed EU Child Sexual Abuse Regulation violate EU (privacy) law,' *European Law Blog*, available at <https://europeanlawblog.eu/2023/05/15/voluntary-detection-orders-under-the-proposed-eu-child-sexual-abuse-regulation-violate-eu-privacy-law/> accessed 28 Aug 2023.

Wolfsfeld, Gadi, Elad Segev, and Tamir Sheafer (2013) 'Social media and the Arab Spring: Politics comes first' *Journal of Press/Politics* 18(2): 115-137.

York, Jillian (2022) *Silicon values: The future of free speech under surveillance capitalism*. Verso Books.

Young, Iris Marion (2011) *Justice And The Politics Of Difference*, Princeton University Press.

Zuboff, Shoshana (2019) *Surveillance capitalism and the challenge of collective action*. New labor forum. Vol. 28. No. 1. Sage CA: Los Angeles, CA: SAGE Publications.

Non-Governmental Organizations Reports

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

Amnesty International (2017), Attacks on human rights activities reach crisis point globally, available at <https://www.amnesty.nl/actueel/attacks-on-human-rights-activists-reach-crisis-point-globally> accessed 7 Sep 2023.

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

Article 19 (2021), At a glance: Does the EU Digital Services Act protect freedom of expression, available at <https://www.article19.org/resources/does-the-digital-services-act-protect-freedom-of-expression/> accessed 10 April 2024.

Article 19 (2021), Europe: Artificial Intelligence Act Must Protect Free Speech and Privacy, ARTICLE19, available at <https://www.article19.org/resources/europe-artificial-intelligence-act-must-protect-freedom-of-expression-and-privacy/> accessed 29 August 2024.

Business and Human Rights Resource Centre (2021), Syria: New report highlights the complicity of multinational tech companies in the regime's human rights violations, available at <https://www.business-humanrights.org/en/latest-news/syria-new-report-highlights-the-complicity-of-multinational-tech-companies-in-the-regimes-human-rights-violations/> accessed 10 April 2024.

Business for Social Responsibility (2022), Human Rights Impact Assessment: Meta's Expansion of End-to-End Encryption, available at <https://www.bsr.org/reports/bsr-meta-human-rights-impact-assessment-e2ee-report.pdf> and available at <https://www.bsr.org/en/reports/metas-expansion-end-to-end-encryption> accessed 10 April 2024.

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

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

European Digital Rights (EDRi) (2014), 'EU Parliament calls for ban of public 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 29 August 2024.

European Digital Rights (EDRi) (2020), "French Avia law declared unconstitutional: what does this teach us at EU level?", available at <https://edri.org/our-work/french-avia-law-declared-unconstitutional-what-does-this-teach-us-at-eu-level/> accessed 18 November 2021.

European Digital Rights (EDRi) (2021), 'How Europol's reform enables 'NSA-style 'surveillance operations', available at <https://edri.org/our-work/how-europols-reform-enables-nsa-style-surveillance-operations/> accessed 17 October 2023.

European Digital Rights (EDRi) (2022), A safe internet for all, Upholding private and secure communication, available at <https://edri.org/wp-content/uploads/2022/10/EDRi-Position-Paper-CSAR.pdf> accessed 7 Sep 2023, 24 and 25.

European Digital Rights (EDRi) (2023) Online Safety Bill insecure: international organisations, academics and cyber experts urge UK government to protect encrypted messaging, available at <https://edri.org/our-work/online-safety-bill-insecure-international-organisations-academics-and-cyber-experts-urge-uk-government-to-protect-encrypted-messaging/> accessed 10 April 2024.

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

Human Rights Watch (2023) Meta's Broken Promises, Systematic Censorship of Palestine Content on Instagram and Facebook, available at <https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and> accessed 10 April 2024.

Tech against terrorism (2021) 'Terrorism use of E2EE: State of Play, Misconceptions, and Mitigation Strategies Report', available at <https://www.techagainstterrorism.org/wp-content/uploads/2021/09/TAT-Terrorist-use-of-E2EE-and-mitigation-strategies-report-.pdf> accessed 28 Aug 2023, 42-56.

Tech Against Terrorism (2021) Submission to the United Kingdom Draft Online Safety Bill Consultation, available at <https://www.techagainstterrorism.org/wp-content/uploads/2021/09/Tech-Against-Terrorisms-Response-%E2%80%93-Joint-Committee-OSB-call-for-written-evidence.pdf> accessed 10 April 2024.

News Articles

ABC News (2023) Donal Trump Supporters embrace Signal, Telegram and other 'free speech' apps, available at <https://www.abc.net.au/news/2021-01-20/donald-trump-social-media-apps-free-speech-privacy/13071206> accessed 7 Sep 2023.

Al Jazeera (2024) 'The Listening Post: Genocide in Gaza: Enabled by AI, powered by Big Tech', available at <https://www.aljazeera.com/program/the-listening-post/2024/4/13/genocide-in-gaza-enabled-by-ai-powered-by-big-tech> accessed 30 May 2024.

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

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

Al Khatib, Hadi and Dia Kayyali (The New York Times, 23 October 2019) 'YouTube Is Erasing History', available at <https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html> accessed 6 April 2023.

Allenbach-Ammann, Janos (EURACTIV, 2023) EU Parliament and member states reach deal on corporate due diligence law, , available at <https://www.euractiv.com/section/economy-jobs/news/eu-parliament-and-member-states-reach-deal-on-corporate-due-diligence-law/> accessed 5 Feb 2024.

Angwin, Julia and Hannes Grassegger (ProPublica, 2027) "Facebook's secret censorship rules protect white men from hate speech but not black children" available at <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms> accessed 10 April 2024.

Angwin, Julia, ProPublica & Hannes Grassegger, Facebook's Secret Censorship Rules Protect White Men From Hate

Arampatzis, Anastasios (2023) Homomorphic Encryption: What Is It and How Is It Used, available at <https://venafi.com/blog/homomorphic-encryption-what-it-and-how-it-used/> accessed 7 Sep 2023.

Arango, Tim, Nicholas Bogel-Burroughs & Katie Benner, Minutes Before El Paso Killing, Hate Filled Manifesto Appears Online, N.Y. TIMES (Aug. 3, 2019), available at <https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html> accessed 29 August 2024.

Asher-Schapiro, Avi & Ban Barkawi (Reuters, 2020) 'Lost Memories': War crimes evidence threatened by AI moderation, available at <https://www.reuters.com/article/us-global-socialmedia-rights-trfn-idUSKB N23Q2TO> accessed 29 August 2024.

Bertuzzi, Luca (2024) EU countries give crucial nod to first-of-a-kind Artificial Intelligence law, EURACTIV, available at <https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/> accessed 5 Feb 2024.

Brodkin, Jon (2022) "War upon end-to-end encryption": EU wants Big tech to scan private messages, available at <https://arstechnica.com/tech-policy/2022/05/war-upon-end-to-end-encryption-eu-wants-big-tech-to-scan-private-messages/> accessed 7 Sep 2023.

Cranz, Alex and Russell Brandom (The Verge, 2021) 'Facebook encourages hate speech for profit, says whistleblower', available at <https://www.theverge.com/2021/10/3/22707860/facebook-whistleblower-leaked-documents-files-regulation>accessed 28 Aug 2023;

Danao, Monique (Forbes, 2023) What can someone do with your IP address? available at <https://www.forbes.com/advisor/business/what-can-someone-do-with-ip-address/#:~:text=IP%20addresses%20can%20be%20used,where%20your%20device%20is%20located.> accessed 5 Feb 2024.

Davies, CJ (The Wired, 2009) The hidden censors of the internet, available at <https://www.wired.co.uk/article/the-hidden-censors-of-the-internet> accessed 5 Feb 2024.

Dearden, Lizzie (Independent, 2018) Gab: Inside the social network where alleged Pittsburgh synagogue shooter posted final message, available at <https://www.independent.co.uk/news/world/americas/pittsburgh-synagogue-shooter-gab-robert-bowers-final-posts-online-comments-a8605721.html> accessed 10 April 2024.

Dearden, Lizzie, (The Independent, 2018) 'Gab: Inside the Social Network Where Alleged Pitts burgh Synagogue Shooter Posted Final Message', (Oct. 28, 2018, 8:10 PM) available at <https://www.independent.co.uk/tech/pittsbur gh-synagogue-shooter-gab-robert-bowers-final-posts-online-comments-a8605721.html> accessed 29 August 2024.

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

Elliot, Larry (The Guardian, 2024) 'Big tech firm recklessly pursuing profits from AI, says UN head', available at <https://www.theguardian.com/business/2024/jan/17/big-tech-firms-ai-un-antonio-guterres-davos> accessed 28 May 2024.

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> accessed 29 August 2024.

Farmer, Ashely D., Organization of American Historians, available at <https://www.oah.org/tah/history-for-black-lives/tracking-activists-the-fbis-surveillance-of-black-women-activists-then-and-now/> accessed 7 Sep 2023.

Foreign Policy (2021) Are Telegram and Signal Havens for Right-Wing Extremists? available at < https://foreignpolicy.com/2021/03/13/telegram-signal-apps-right-wing-extremism-islamic-state-terrorism-violence-europol-encrypted/#cookie_message_anchor> accessed 7 Sep 2023.

Fornetix (2022) End-to-End Social Media Encryption Strategies, available at <https://www.fornetix.com/articles/end-to-end-encryption-strategies-becoming-the-norm-for-social-media/> accessed 7 Sep 2023;

Ganster, Allyson M 'Black women and digital resistance: The impact of social media on racial justice activism in Brazil and the United States' Diss. 2019, available at <https://repositories.lib.utexas.edu/items/45168a42-b43d-47ea-800f-24cd7d2d04cc> accessed 28 May 2024.

Giansiracusa, Noah (Wired, 2021) 'Facebook Uses Deceptive Match to Hide its Hate Speech Problem', available at <https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/> accessed 17 October 2023.

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

Hao, Karen (MIT Technology Review, 2020) We read the paper that forced Timnit Gebru out of Google. Here's what is says. available at <https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/> accessed 10 April 2024.

Hao, Karen (MIT Technology Review, 2021) 'The Facebook whistleblower says its algorithms are dangerous. Here's why.', available at <https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/> accessed 28 Aug 2023

Harper, Shaun (Forbes, 31 October 2022) 'Hate Speech Rises On Twitter After Elon Musk Takes Over, Researchers Find', 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.

Kanu, Hassan (Reuters, 2022) Prevalence of white supremacists in law enforcement demands drastic change, available at <https://www.reuters.com/legal/government/prevalence-white-supremacists-law-enforcement-demands-drastic-change-2022-05-12/> accessed 7 Feb 2024.

Kleinman, Zoe and Tom Gerken (2023) Twitter launches encrypted private messages, says Elon Musk, available at <https://www.bbc.com/news/technology-65533021> accessed 7 Sep 2023.

Koomen, Maria (Carnegie Endowment for International Peace, 2021) 'The Encryption Debate in the European Union: 2021 Update', available at <https://carnegieendowment.org/2021/03/31/encryption-debate-in-european-union-2021-update-pub-84217> accessed 28 Aug 2023.

Layug, Alyan, et al. 'The impacts of social media use and online racial discrimination on Asian American mental health: cross-sectional survey in the United States during COVID-19." JMIR formative research 6.9 (2022): e38589 available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9488547/> accessed 28 May 2024.

Lunden, Ingrid (2020) Facebook adds hosting, shopping features and pricing tiers to WhatsApp Business, available at <https://rb.gy/2sj7p> accessed 7 Sep 2023.

Lutkevich, Ben and Madelyn Bacon (2021) Definition end-to-end encryption (E2EE) available at <https://www.techtarget.com/searchsecurity/definition/end-to-end-encryption-E2EE> accessed 7 Sep 2023.

Medianama, Sarvesh Methi (2022) "How end-to-end encryption impact human rights? The Good and the Bad" available at <https://www.medianama.com/2022/04/223-end-to-end-encryption-human-rights-impact/> accessed 10 April 2024.

O'Flaherty, Kate (Wired, 2018) YouTube Keep deleting evidence of Syrian chemical weapon attacks, available at <https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video> accessed 7 Sep 2023.

OT, Anina (2021) What Apps Use End-to-End Encryption to Improve Online Privacy, available at <https://www.makeuseof.com/apps-use-end-to-end-encryption/> accessed 7 Sep 2023.

Prakken d'Oliveira (2018) European Court: Decisions Placing the PKK on the List of Terrorist Organizations Annulled, Prakken d'Oliveira (Nov. 15, 2018) available at <https://www.prakkendoliveira.nl/en/news/2018/european-court-decisions-placing-the-pkk-on-the-list-of-terrorist-organizations-annulled> accessed 29 August 2024.

Purnell, Newley and Jeff Horwitz, 'Facebook Services Are Used to Spread Religious Hatred in India, Internal Documents.

Rachwani, Motafa and Christopher Knaus (The Guardian, 2023) Videos urged counter-protesters to attack LGBTQ+ activists outside Sydney church, available at <https://www.theguardian.com/australia-news/2023/mar/22/videos-urged-counter-protesters-to-attack-lgbtq-activists-outside-sydney-church> accessed 7 Sep 2023.

Ray, Siladitya (2023) Encrypted Messaging, 2-Hour Videos: Here Are the Moves Twitter Has Made in Its Bid To Become an 'Everything' App, available at <https://www.forbes.com/sites/siladityaray/2023/05/26/encrypted-messaging-2-hour-

videos-here-are-the-moves-twitter-has-made-in-its-bid-to-become-an-everything-app/> accessed 7 Sep 2023.

Robertson, Adi (the Verge, 2021) 'Apple's controversial new child protection features, explained', available at <https://www.theverge.com/2021/8/10/22613225/apple-csam-scanning-messages-child-safety-features-privacy-controversy-explained> accessed 7 Sep 2023.

Safi, Michael (The Guardian, 2018) Sri Lanka accuses Facebook over hate speech after deadly riots, available at <https://www.theguardian.com/world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech> accessed 10 April 204.

Safi, Michael, Sri Lanka Accuses Facebook Over Hate Speech After Deadly Riots, (Guardian, 2018), available at <https://www.theguardian.com/ world/2018/mar/14/facebook-accused-by-sri-lanka-of-failing-to-control-hate-speech> accessed 29 August 2024.

Show' (The Wall Street Journal, 2021) available at <https://www.wsj.com/articles/facebook-services-are-used-to-spread-religious-hatred-in-india-internal-documents-show-11635016354?mod=article_inline> accessed 28 Aug 2023.

Smith, Ashel (Bits of Freedom, 2022) European Commission wants to eliminate online confidentiality, available at <https://www.bitsoffreedom.nl/2022/05/11/european-commission-wants-to-eliminate-online-confidentiality/> accessed 7 Sep 2023;

Spadafora, Anthony (2023) The best encrypted messaging apps in 2023 available at <https://www.tomsguide.com/reference/best-encrypted-messaging-apps> accessed 7 Sep 2023.

Speech But Not Black Children, (PROPUBLICA, 2017), available at <https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms> accessed 21 Feb 2024.

Stevenson, Alexandra (New York Times, 2018), 'Facebook Admits It Was Used to Incite Violence in Myanmar', available at <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html> accessed 29 August 2024.

Sumner, A. et al. (Jama Network Open, 2021) Association of Online Risk Factors with Subsequent Youth Suicide-Related Behaviors in the US, available at <https://jamanetwork.com/ journals/jamanetworkop en/fullarticle/2784337Steven> accessed 29 August 2024.

The New York Times (2018) Facebooks admits it was used to incite violence in Myanmar, available at <https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html> accessed 10 April 2024.

The New York Times (2019) Minutes before El Paso killing, hate filled manifesto appears online, available at <https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html?action=click&module=Spotlight&pgtype=Homepage> accessed 10 April 2024.

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.

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 29 August 2024.

The Wall Street Journal (2020) 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive', available at <wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499> accessed 28 May 2024.

The Wall Street Journal (2021) The Facebook Papers, available at <https://facebookpapers.com/outlet/wall-street-journal/> accessed 10 April 2024.

UCD Centre for Digital Policy, available at <https://digitalpolicy.ie/explainer-irelandsonline-safety-and-media-regulation-bill/> accessed 18 November 2021.

Vincent, James (The Verge, 2022), New EU riles would require chat apps to scan private message for child abuse, available at <https://www.theverge.com/2022/5/11/23066683/eu-child-abuse-grooming-scanning-messaging-apps-break-encryption-fears?scrolla=5eb6d68b7fedc32c19ef33b4> accessed 7 Sep 2023.

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

Whitney, Lance (2023) "Twitter rolls out encryption for direct messages but with key limitations" available at <https://www.zdnet.com/article/twitter-rolls-out-encryption-for-direct-messages-but-with-key-limitations/> accessed 10 April 2024.

Wikipedia, FBI requested backdoors to Apple's iPhone software, available at <https://en.wikipedia.org/wiki/End-to-end_encryption#Backdoors> accessed 7 Sep 2023.

Williams, Adrienne, Milagros Miceli and Timnit Gebru 'The Exploited Labor Behing Artificial Intelligence' (2022) available at <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/> accessed 28 May 2024.

Zenger, Rejo (Bits of Freedom, 2022) This might sound attention-seeking, but we really believe to be not far off the mark. It really looks like the European Commission wants to cancel encryption, available at <https://www.bitsoffreedom.nl/2022/05/11/european-commission-wants-to-eliminate-online-confidentiality/> accessed 28 August 2024.

Other Web Resources

Content Policy, Reddit, available at <Https://Www.Redditinc.Com/Policies/Content-Policy#:~:Text=Abide%20by%20community%20rules.,With%20or%20Disrupt%20Reddit%20communities.&Amp;Text=Respect%20the%20privacy%20of%20others> Accessed 29 August 2024.

Daniel Ruby, '55+ Facebook Statistics For 2023 (Users, Revenue & Trends)' (Demand Sage, 10 February 2023) available at <Https://Www.Demandsage.Com/Facebook-Statistics/> Accessed 29 August 2024.

Emily R (2022) Top 7 Most Secure Video Calling Apps, available at <Https://Getstream.Io/Blog/Safest-Video-Calling-Apps/> Accessed 7 Sep 2023.

Facebook Community Standards Hate Speech (2023) available at <Https://Transparency.Fb.Com/Pt-Pt/Policies/Community-Standards/Hate-Speech/> Accessed 29 August 2024.

Facebook Help Center, What End-To-End Encryption On Messenger Means And How It Works, available at <Https://Www.Facebook.Com/Help/Messenger-App/786613221989782?Cms_Id=786613221989782> Accessed 21 February 2024.

Facebook Help Centre, available at <Https://Www.Facebook.Com/Help/Messenger-App/786613221989782?Cms_Id=786613221989782> Accessed 28 Aug 2023.

Google, 'Featured Policies: Hate Speech' (2023) available at <Https://Transparency report.Google.Com/Youtube-Policy/Featured-Policies/Hate-Speech?Hl=En> Accessed 6 April 2023.

How Meta Enforces Its Policies (2022) available at <Https://Transparency.Fb.Com/En-Gb/Enforcement/> Accessed 29 August 2024.

How Meta Prioritises Content For Review (2022) available at <Https://Transparency.Fb.Com/Policies/Improving/Prioritizing-Content-Review/> Accessed 29 August 2024.

Idowu Omisola (2023) WhatsApp Community Vs. WhatsApp Group: What's The Difference? available at <Https://Www.Makeuseof.Com/Whatsapp-Community-Vs-Whatsapp-Group-Difference/> Accessed 7 Sep 2023.

Instagram Help Centre (2023) How Do I Start An End-To-End Encrypted Chat On Instagram, available at <Https://Help.Instagram.Com/1165835007222763/?Helpref=Related_Articles> Accessed 7 Sep 2023.

Janice Asare, 'Are Marginalized Communities Being Censored Online' (2020) Forbes, available at <Https://Www.Forbes.Com/Sites/Janicegassam/2020/05/24/Are-Marginalized-Communities-Being-Censored-Online/> Accessed 28 May 2024.

Jim Waterson & Dan Milmo, Facebook Whistleblower Frances Haugen Calls For Urgent External Regulation, Guardian, available at <https://www.theguardian.com/technology/2021/oct/25/facebook-whistleblower-frances-haugen-calls-for-urgent-external-regulation> accessed 28 November 2024.

Jim Waterson And Dan Milmo, Facebook Whistleblower Frances Haugen Calls For Urgent External Regulation (Oct. 25, 2021) available at <Https://www.Theguardian.Com/Technology/2021/Oct/25/Facebook-Whistleblower-Frances-Haugen-Calls-For-Urgent-External-Regulation> Accessed 29 August 2024.

LinkedIn, Help ""Hateful And Derogatory Content" available at <Https://Www.Linkedin.Com/Help/Linkedin/Answer/A1339812> Accessed 11 August 2024.

Mansoor Iqbal (2023) Facebook Revenue And Usage Statistics, available at <Https://Www.Businessofapps.Com/Data/Facebook-Statistics/> Accessed 7 Sep 2023.

Mansoor Iqbal (2023) Telegram Revenue And Usage Statistics, available at <Https://Www.Businessofapps.Com/Data/Telegram-Statistics/> Accessed 7 Sep 2023.

Mansoor Iqbal (2023) WhatsApp Revenue And Usage Statistics, available at <Https://Www.Businessofapps.Com/Data/Whatsapp-Statistics/> Accessed 7 Sep 2023.

Mansoor Iqbal, 'Twitter Revenue And Usage Statistics (2023)' (Business Of Apps, 21 January 2023) available at <Https://Www.Businessofapps.Com/Data/Twitter-Statistics/> Accessed 29 August 2024.

Meta (2019) Detecting Non-Consensual Intimate Images And Supporting Victims, available at <Https://About.Fb.Com/News/2019/03/Detecting-Non-Consensual-Intimate-Images/> Accessed 7 Sep 2023.

Meta (2023) Messenger End-To-End Encryption Overview, available at <Https://Engineering.Fb.Com/Wp-Content/Uploads/2023/12/Messengerend-To-Endencryptionoverview_12-6-2023.Pdf> Accessed 10 April 2024.

Meta, Transparency Center "Hate Speech" available at <Https://Transparency.Meta.Com/En-Gb/Policies/Community-Standards/Hate-Speech/> Accessed 11 August 2024.

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.

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 29 August 2024.

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 6 April 2023.

Professional Community Policies, LinkedIn, available at <Https://Www.Linkedin.Com/Legal/Professional-Community-Policies> Accessed 29 August 2024.

Signal Support, available at <Https://Support.Signal.Org/Hc/En-Us/Articles/360007320391-Is-It-Private-Can-I-Trust-It> Accessed 10 April 2024.

Signal Support, Group Chats, available at <Https://Support.Signal.Org/Hc/En-Us/Articles/360007319331-Group-Chats#:~:Text=Admin%20controls%20of%20who%20can%20send%20messages%20and%20start%20calls,Size%20limit%20of%201000> Accessed 7 Sep 2023.

Similarweb, Top Websites Ranking (2023) available at <Https://Www.Similarweb.Com/Top-Websites/> Accessed 6 April 2023.

Snow, D. Della Porta, D., Klandermans, B. And Mcadam, D. (Eds.) Encyclopedia Of Social And Political Movements, Agents Provocateus As A Type Of Faux Activist, available at <Https://Web.Mit.Edu/Gtmarx/Www/Agentsprovocateursfaux.Html> Accessed 7 Sep 2023.

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.

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.

Statistica (2024), Worldwide Digital Population, available at <Https://Www.Statista.Com/Statistics/617136/Digital-Population-Worldwide/#:~:Text=Worldwide%20digital%20population%202024&Text=As%20of%20January%202024%2C%20there,Population%2C%20were%20social%20media%20users> Accessed 10 April 2024.

Statistica, Social Media – Statistics And Facts, available at <Https://Www.Statista.Com/Topics/1164/Social-Networks/#Topicoverview> Accessed 21 Feb 2024.

Telegram FAQ, available at <Https://Telegram.Org/Faq#Secret-Chats> Accessed 10 April 2024.

Telegram Group Chats On Telegram, available at <Https://Telegram.Org/Faq#:~:Text=With%20Telegram%2C%20you%20can%20send,For%20broadcasting%20to%20unlimited%20audiences> Accessed 7 Sep 2023.

Timothy Buck (2022) Update To End-To-End Encrypted Chats On Messenger, available at <Https://About.Fb.Com/News/2022/01/Updates-To-End-To-End-Encrypted-Chats-Messenger/> Accessed 7 Sep 2023.

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.

Twitter Help Center, 'Hateful Conduct' (2023) available at <Https://Help.Twitter.Com/En/Rules-And-Policies/Hateful-Conduct-Policy> Accessed 6 April 2023.

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.

WhatsApp Help Center – How WhatsApp Helps Fight Child Exploitation. available at <Https://Faq.Whatsapp.Com/General/How-Whatsapp-Helps-Fight-Child-Exploitation/?Lang=En> Accessed 7 Sep 2023.

WhatsApp Help Center, About End-To-End Encryption, available at <Https://Faq.Whatsapp.Com/820124435853543?Locale=En_US&Cms_Id=820124435853543&Draft=False> Accessed 10 April 2024.

WhatsApp Help Center, How To Add And Remove Group Participants, available at <Https://Faq.Whatsapp.Com/841426356990637/?Locale=En_US&Cms_Platform=Web&Cms_Id=841426356990637&Draft=False> Accessed 7 Sep 2023.

WhatsApp Terms Of Service, WhatsApp, available at <Https://Www.Whats App.Com/Legal/Terms-Of-Service/?Lang=En> Accessed 29 August 2024.

X, Help Center "Hateful Conduct" available at < Https://Help.X.Com/En/Rules-And-Policies/Hateful-Conduct-Policy> Accessed 11 August 2024.

YouTube Help, 'Reporting And Enforcement' (2023) available at <Https://Support.Google.Com/Youtube/Topic/2803138?Hl=En&Ref_Topic=6151248> Accessed 6 April 2023.

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/Howyoutube works/Our-Commitments/Standing-Up-To-Hate/> Accessed 6 April 2023.

Audio Resources

· Podcast series by Katie Pentney "Decoding Hate," Episode 2 "The Hate You Tweet" with Tarlach McGonagle, 10 February 2021, funded by OSCE Representative on Freedom of the Media, #SAIFE project, available at <https://www.decodinghatepod.com/episodes/episode-05-the-anywhere-.workout-lhgdz-D0JBu> accessed 10 April 2024.

List of publications and talks

Ph.D. Publications

1 *Nave E.* (2023), Hate speech, historical oppressions, and European human rights, *Buffalo Human Rights Law Review* 29 (2022/2023): 83-145, https://digitalcommons.law.buffalo.edu/bhrlr/vol29/iss1/3/.

2 *Nave E.* and Lane L. (2023), Countering online hate speech: how does human rights due diligence impact terms of service?, *Computer Law and Security Review* 51: 105884, https://doi.org/10.1016/j.clsr.2023.105884.
-- *Nave E.* and Lane L. (2024) Governing hate speech detection on online platforms: A human rights approach. *International Joint Conference on Artificial Intelligence 2024 Workshop on AI Governance: Alignment, Morality, and Law*.

3 *Nave E.,* Raaijmakers S., and Veugen T (2024) Disrupting violence while preserving encryption: A human rights approach, *Technology and Regulation* 2024 (2024): 115-131, https://doi.org/10.26116/techreg.2024.012.

4 *Nave E.* (under review, 2024) Criminal hate speech attributable to online platforms – a call for a thorough corporate remedial responsibilities framework in Europe.

Other selected publications

Tan, Y., *Nave, E.,* Vandebosch, H., Pabian, S., Poels, K. (2023) Methods for reporting online sexual harassment (NETHATE report), https://doi.org/10.31235/osf.io/p92t4.

Nave, E. (2019) The Importance of the Arms Trade Treaty for the Implementation of the Sustainable Development Goals. Journal of Conflict and Security Law, 24(2), 297-324 (Oxford Press): https://doi.org/10.1093/jcsl/krz010.

Brown, M., Burkard, C., Jeangeorge, A., *Nave, E.* (2017) Computer Modelling in Collateral Damage Estimates and Choice of Weapons in Iraq. ICRC International Humanitarian Law in Action: respect for the law on the battlefield, supervised by Heinsch, R.W., Poulopoulou, S., and Tremblay, C; 6 other co-authored case-studies of Colombia, Eritrea/Ethiopia, Iraq and Peru: https://ihl-in-action.icrc.org/.

Nave, E. (2016) Can drones comply with the IHL targeting framework? Critique about drones and compliance with the principle of distinction, Research

paper, Privatissimum in Contemporary Issues in International Humanitarian Law.

Nave, E. (2016) What would I do if I was Abbas, Netanyahu or myself to improve the peace process between the Israeli and the Palestinians? Youth Peace Initiative International Essay Competition (awarded 2nd prize).

Nave, E. (2010) European Citizenship, a sovereignties' monologue?, Research paper for the European Union Law course, available at the NOVA Law School online repository (best class paper).

Selected Talks resulting from the Ph.D. research

Invited speaker at the International Network for Hate Studies, 2024 Biennial Conference, NETHATE panel "Understanding Online Hate" (25-27/11/2024).

Guest lecture titled "Hate speech, human rights and content moderation", 2 hours guest lecture for the?Pre-University College course on Law and Technology,?eLaw – Center for Law and Digital Technologies, Leiden University (Leiden, The Netherlands, 26/02/2024 and 15/02/2023).

Invited talk titled "Countering online hate speech: How does human rights due diligence impact terms of service?", European seminar series on Legal Issues Arising from Networks, https://www.lians-seminar.com/speakers, joint work with Dr. Lottie Lane (online, 18/10/2023).

Speaker in the panel titled "Content moderation at the infrastructure level and its impact on business and human rights", organized by ARTICLE 19 at the RightsCon 2023 Edition (online, 07/06/2023).

Guest lecture "Hate speech, human rights and content moderation", 2 hours, Fundamental Rights course, Bachelor's Programme Social Sciences for Globalization, UNIMI, Italy (online, 09/05/2023).

Invited talk titled "Hate speech, historical oppressions, and European human rights", "International Seminar Hate Speech: How to Counteract?", University of Azores, Portugal (online, 04/05/2023).

Invited talk titled "Countering online hate speech: How does human rights due diligence impact terms of service?", Laborattori ittineranti di Diritto internazionale, UNIMI (Milan, Italy, 13/04/2023).

Invited talk titled "Countering online hate speech in Europe through enhanced human rights due diligence responsibilities in the drafting of terms of service", "Digital Legal Talks 2022 Conference", joint work with Dr. Lottie Lane (Utrecht, The Netherlands, 24/11/2022).

Speaker/ participant in the International Workshop "Hate Speech – an Interdisciplinary Approach", organized by the Minerva Center for the Rule of Law under Extreme Conditions at the University of Haifa in collaboration with Freie Universitat Berlin and Technischen Universitat Berlin (online, 17-19/01/2022).

Speaker/ participant in the Workshop: "Method, methodology and critique in international law" (online, 15-18 December 2021).

Invited speaker in the Podcast series "Integrated Projects" created by students at the Leiden University's Security Studies Bachelor's degree, episode on "Hate speech, cyberbullying and doxing" (online, 2021).

Invited talk titled "Countering online hate speech – how to effectively protect fundamental rights?", Annual Conference of the research programme "Effective Protection of Fundamental Rights in a pluralist word", edition titled "Legal regimes and social divides – radicalization, discrimination and exclusion" (online, 2021).

Invited speaker in the Fourth Annual Netherlands Network for Human Rights Research (NNHRR) Conferences "Human Rights and Vulnerability", participation in the panel on "Vulnerability and Human Rights in the Digital Age" convened by the Human Rights in the Digital Age Working Group (online, 2021).

Curriculum vitae

Eva Nave was born in Guarda, Portugal, in 1991. Eva holds a LL.B. in Law at the Nova University of Lisbon, with one semester completed at LMU Munich, a Postgraduation in Human Rights at the Coimbra University, and a LL.M. in Public International Law at the Leiden University. During her LL.M., she was part of the International Humanitarian Law Clinic where she integrated as junior researcher an international team working for the International Committee of the Red Cross.

Since 2021, Eva is a PhD candidate at eLaw – Center for Law and Digital Technologies, Leiden University and Marie Curie Fellow with the Marie Sk³odowska-Curie Actions Innovative Training Network project titled Network of Excellence for Training on Hate (NETHATE). Her research focuses on countering hate speech though enhanced corporate human rights due diligence for online platforms and, in 2024, Eva was awarded the Leiden University prize of best 2022-2023 Ph.D. paper on human rights. During her Ph.D., Eva was a visiting researcher at the University of Queensland (05-06/24), under the mentorship of Professor Katharine Gelber, and at the University of Milan UNIMI (03-04/23), and a seconded Researcher at TNO Netherlands Organization for Applied Scientific Research (10/22-01/23). From June 2024 to April 2025, Eva was also an Advisor to the Portuguese Secretary of State for Science, on legal and policy matters related to human rights, public administration for science, and law and technology.

Before joining eLaw, Eva worked for four years with the United Nations Mine Action Service, where she contributed with legal research, policy and project advice on weapons control and conflict prevention in Mali, the Democratic Republic of the Congo, Uganda and Colombia. Prior to that, she worked for two years in the human rights field both with non-governmental organizations (Amnesty International, the Portuguese Refugee Council and the Portuguese Association for Victims Support) and with governmental bodies (the Boston Mayor's Office for Immigrants Advancement). Eva was also a trainee at the Secretariat for the Committee on Constitutional Affairs at the European Parliament.

In the range of books published by the Meijers Research Institute and Graduate School of Leiden Law School, Leiden University, the following titles were published in 2024-2025:

MI-414 E. Hutten, *Belastingprofessionals onder maatschappelijke druk. Een Nederlandse casestudie naar reacties op BEPS*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-415 R. Stolk, *Procederende belangenorganisatie in de polder. Een interdisciplinair perspectief op de toegang tot de rechter*, (diss. Leiden), Zutphen: Uitgeverij Paris 2024

MI-416 A. Sarris, *International law and governance of the Arctic in an era of climate change*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024, ISBN 978 94 6473 382 2

MI-417 F. Heitmüller, *Combatting tax avoidance, the OECD way? The impact of the BEPS Project on developing and emerging countries' approach to international tax avoidance,* (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-418 F.I. Kartikasari, *Mining and environmental protection in Indonesia: Regulatory pitfalls,* (diss. Leiden), Amsterdam: Ipskamp Printing 2024, ISBN 978 94 6473 462 1

MI-419 S.H. Starrenburg, *Striking a balance between local and global interests. Communities and cultural heritage protection in public international law*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-420 D. Stefoudi, *Legal and policy aspects of space big data. Legal implications of the use of large amounts of space data – Regulatory solutions and policy recommendations* (diss. Leiden), Amsterdam: Ipskamp Printing 2024, ISBN 978 94 6473 479 9

MI-421 S. Poulopoulou, *Towards the establishment of a new International Humanitarian Law compliance mechanism. Lessons learned from monitoring systems within the International Humanitarian and Human Rights Law frameworks*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-422 M. Aalbers, *De werking van algemene belangenafwegingen in het Europese staatssteunrecht. Tussen verbod en verenigbaarheid?*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-423 J.M. Elbers, *Reward Systems in Prison*, (diss. Leiden), Alblasserdam: Ridderprint 2024

MI-424 Z. Tian, *Legal Aspects of Active Debris Removal (ADR): Regulation of ADR under International Space Law and the Way Forward for Legal Development*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-425 J.P. Cnossen, *Wisselwerking tussen commuun en bijzonder materieel strafrecht. Een analyse en waardering in het licht van de beginselen van codificatie, schuld en legaliteit*, (diss. Leiden), Den Haag: Boom juridisch 2024, ISBN 978 94 6212 967 2, ISBN 978 94 0011 466 1 (e-book)

MI-426 L.B. Louis, *Towards Better Policing. Achieving Norm Internalization and Compliance with Persuasively Designed Technology,* (diss. Leiden), Amsterdam: Ipskamp Printing 2024, ISBN 978 94 6473 559 8

MI-427 J.P. Loof & R.A. Lawson (red.), *Diverse mensen en gelijke rechten anno 2024. Essays ter gelegenheid van het emeritaat van prof. Titia Loenen als hoogleraar Mensenrechten en diversiteit*, Leiden: Stichting NJCM-Boekerij 2024, ISBN 978 90 6750 070 8

MI-428 Y. Shi, *Labour Regulation of International Aviation. A Crawl-Walk-Run Approach in International Law*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024, ISBN 978 94 6473 588 8

MI-429 I.S. Ouwehand, *Toetsing van deskundigenadviezen door de bestuursrechter*, (diss. Leiden), Zutphen: Uitgeverij Paris 2024, ISBN 978 94 6251 362 4

MI-430 P.L. Koopmans, *Essays on the Economics of Household Finance and Social Insurance*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-431 N.U. van Capelleveen, *Radicalisering bij minderjarigen en overheidsingrijpen. Over de interactie van rechtsgebieden en een kinder- en mensenrechtenconforme inzet van juridische instrumenten*, (diss. Leiden), Den Haag: Boom 2024, ISBN 978 94 6212 009 9, ISBN 978 94 0011 504 0 (e-book)

MI-432 A.B. Muñoz Mosquera, *The North Atlantic Treaty Organization. An International Institutional Law Perspectivee*, (diss. Leiden), Amsterdam: Ipskamp Printing 2024

MI-433 S. Vandenbroucke, *Navigating Corporate Responsibility in Global Supply Chains using Codes of Conduct*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-434 B.N. van Ganzen, *Dynamism and Democracy. Essays on the Fiscal Social Contract in a Globalised World*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-435 M. Michels, *Meerouderschap en het erfrecht. Een onderzoek naar de erfrechtelijke positie van het kind en zijn ouders in een intentioneel meeroudergezin*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-436 D.K. Jongkind, *Netwerksubsidies. Een onderzoek naar de wijze waarop samenwerking in subsidierelaties binnen het bestuursrecht kan worden vormgegeven*, (diss. Leiden), Deventer: Kluwer 2025, ISBN 978 90 1318 051 0

MI-437 G. Boffi, *Socio-Economic Integration and Social Citizenship of Migrants: Empirical Analyses*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-438 A. Kaviani Johnson, *From concept to application: A critical reflection on child safeguarding from a children's rights perspective*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025 ISBN 978 94 6473 734 9

MI-439 J. Choi, *Criminal Liability of Pilots in Aviation Accident Cases,* (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-440 K. Sharma, *The Assembly of States Parties to the International Criminal Court – A Good Governance Approach,* (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-441 B. Budinská, *The European Central Bank's centralised application of national law under the Single Supervisory Mechanism. A rule of law analysis*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-442 H. Bliersbach, *Becoming and Belonging? Lived Experiences of Naturalization and Implementation of Citizenship Law in Germany and Canada*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-443 D.P.L. van Thiel, *Fundamental Labour Standards and the Shift from International to Transnational Labour Law. Countervailing Power in the Globalised World of Work*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025, ISBN 978 94 6473 757 8

MI-444 L.M.J. van Doorn, *From Risks to Public Opinion. How Structural Economic Changes Shape Political Attitudes and Policy Preferences*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-445 S.D. Koning, *De ventielfunctie van de artikel 12 Sv-procedure. Klachten tegen niet-vervolging in maatschappelijk gevoelige kwesties*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025, ISBN 978 94 6473 796 7

MI-446 A.S. Florescu, *Migration, Abduction and Children's Rights. The relevance of children's rights and the European supranational system to child abduction cases with immigration components,* (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-447 C. Smit, *Minderheidskabinetten in Nederland en Denemarken*, (diss. Leiden), Den Haag: Boom juridisch 2025, ISBN 978 94 6212 216 1 (978-94-0011-571-2 ebook)

MI-448 S.J. Lopik, *Klimaatstrafrecht. De rol van het strafrecht binnen het juridische antwoord op klimaatverandering*, (diss. Leiden), Deventer: Kluwer 2025, ISBN 978 90 1318 157 9

MI-449 E. Grosfeld, *Increasing Public Perceived Legitimacy of the European Union Through the Integration of Psychological Insights into Law,* (diss. Leiden), Amsterdam: Ipskamp Printing 2025

MI-450 E. Nave, *Countering online hate speech. How to adequately protect fundamental rights?*, (diss. Leiden), Amsterdam: Ipskamp Printing 2025 ISBN 978 94 6473 832 2