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

Nave, E.V.R.

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.

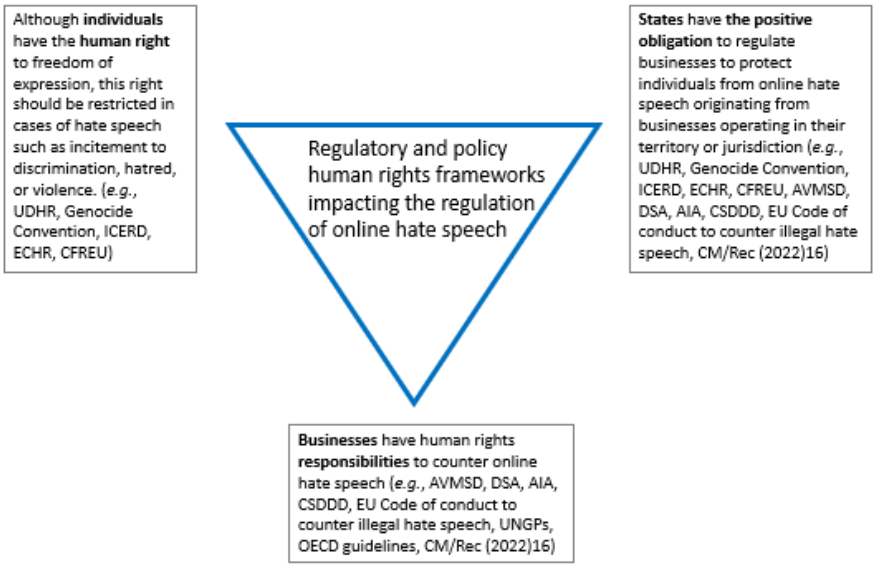

Figure 1 – Overview of regulatory and policy human rights frameworks impacting the regulation of online hate speech.

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.