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

Nave, E.V.R.

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.

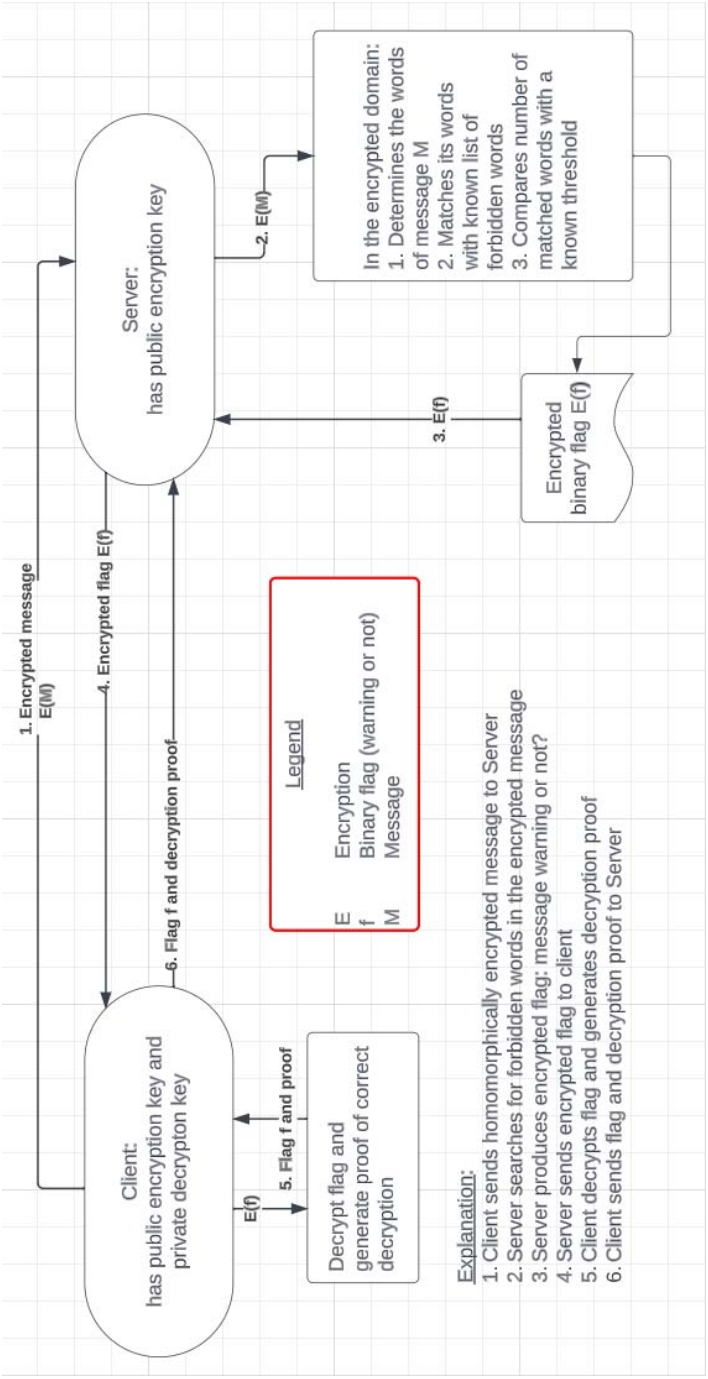

Figure 4 – Homomorphic approach to secure message analysis.

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.