Queen Mary
University of London

Regulating Hate Speech: EU Intermediary Liability Policy and Online Platforms' Tools

Word Count:9953

Supervisor: Dimitra Kamarinou

Student Number:190527783

Technology, Media and Telecommunications Law LLM

8/17/20

Table of Contents

Acknowledgements

This document has been produced with the financial assistance of the Ministry of National Education of Turkey. The contents of this document are the sole responsibility of 190527783 and can under no circumstances be regarded as reflecting the position of Turkey.!

Regulating Hate Speech: EU Intermediary Liability Policy and Online Platforms' Tools

Introduction

> "Facebook has been a useful instrument for those seeking to spread hate, in a context where, for most users, Facebook is the Internet."[1]

Online Platforms offer us new possibilities for using fundamental rights by means of promoting democratic discussion.[2] Yet there is no standardised concept/definition of the term "Online Platform", so experts or politicians decide on definition according to context.[3] Online platforms such as web-hosting providers or internet service providers, are intermediaries. This essay will follow the European Commission approach, which[4] provides a general overview of the most common features of Online Platforms. Firstly, the Online Platform (OP) has a networking feature whereby each new user can potentially enhance all current users' experiences. Secondly, Online Platforms have the capacity to enable immediate user-to-user connections and interest creation. Finally, Online Platforms have the capacity to capture, use and analyse vast amounts of personal and non-personal data to enhance, among other things, the user's content and experience.[5]

The existence of public forums is necessary for free speech.[6] Online networks are important to us as a forum[7] for free speech and can be considered as a "modern public square."[8] However,

[1] Human Rights Council, 'Report of The Independent International Fact-Finding Mission on Myanmar' (2018) UN Doc A/HRC/39/64.
[2] Judit Bayer, 'Between Anarchy and Censorship' (2019) 3 CEPS Papers in Liberty and Security in Europe 34, 1.
[3] Laura Rozenfeldova and Pavol Sokol, 'Liability Regime of Online Platforms New Approaches and Perspectives' (2019) 3 EU and Comparative Law Issues and Challenges Series 866, 868.
[4] European Commission, 'Online Platforms and the Digital Single Market. Opportunities and Challenges for Europe' (2016) Communication from the Commission COM(2016) 0288 final.
[5] ibid 3.
[6] Frederick Mostert, 'Free Speech and Internet Regulation' (2019) 14 Journal of Intellectual Property Law & Practice 607, 607.
[7] Great Britain and others, *Online Harms White Paper* (2019).
[8] *Packingham v North Carolina* [2017] Supreme Court of the United States 137 S. Ct. 1730 (2017) 198 L. Ed. 2d 273.

these sharing opportunities, as offered by Online Platforms, have also opened the door to a broader[9] spectrum of offensive content,[10] and in this new era, the data for online hate speech is greater than "real-life" hate speech.[11] People can quickly to sign in to Online Platforms to express their thoughts to the world without any expense or, indeed, inhibition[12] while essentially being able to maintain their anonymity. Although online platforms do not force people to do hate speech or violence, online platforms provide such people with an echo forum that reinforced them and promoted their offensive views.[13]

However a few decades ago, governments or affluent notables typically dominated mainstream media outlets. Generally, just a few people possessed sufficient influence to convey their opinions through the mass media.[14] Theoretically, the editorial editing in mass media ensures that the content is not discriminatory and hateful, in contrast to, Online Platforms which have been operated without such a filter from the outset, allowing the uncheched spread of such illegal content. The prevalence of hate speech in the context of xenophobic, nationalistic, Islamophobic, racist and anti-Semitic content in online discourse has risen dramatically in recent years.[15] This problem given rise to new questions, such as who is liable for such illegal content? Who will moderate in which settings? One of the steps taken to deal with hate speech and to answer the above questions has engaged the interest of numerous stakeholders such as

[9] Allyson Haynes Stuart, 'Social Media, Manipulation, and Violence' (2019) 15 South Carolina Journal of International Law and Business 100, 101.
[10] Federica Casarosa, 'The European Regulatory Approach toward Hate Speech Online: The Balance between Efficient and Effective Protection' (2020) 22 Gonzaga Journal of International Law 391, 393.
[11] Chinmayi Arun, 'Making Choices: Social Media Platforms and Freedom of Expression Norms' [2018] SSRN Electronic Journal 1 <https://www.ssrn.com/abstract=3411878> accessed 6 July 2020.
[12] Petra Bard and Judit Bayer, 'Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches' (2020).
[13] Stuart (n 9) 118.
[14] Zi En Chow, 'Evaluating the Approaches to Social Media Liability for Prohibited Speech.' (2019) 51 New York University Journal of International Law and Politics 1293, 1298.
[15] European Commission and others, *Media Pluralism and Democracy: Report.* (2016) <http://dx.publications.europa.eu/10.2838/248670> accessed 29 July 2020.

content creators, online platforms and governments, because, as mentioned above, contact with social media is already an integral part of the public's perception of free speech.[16]

The first chapter of this paper will examine hate speech and briefly, its legal history. Government liability approaches will be argued in this section in order to understand Online Platforms' responsibilities and rights too. This paper will defend the moderate liability system which is followed by the European Union, because as will be examined below, in strictly liability systems, Online Platforms may delete entirely legal contents, without the balanced protection of the freedom of speech, purely to reduce the possibility of getting prosecuted.[17]

The E-Commerce Directive[18] (ECD) does not itself contain the notion of an online platform. Still, it includes the notion of 'information society services – referred to as online intermediaries' which host and transmit third-party content. One of the key concerns related to the ECD is to what degree the latest online services, after the implementation of the Directive, fit under the scope of the "information society service"[19] which is a prerequisite to any associated immunity from liability.[20] This essay will accept online platforms as information society services. The EU legal framework on content moderation has become more complicated through categorising Online Platforms and through its risk-based approach.[21] The E-Commerce Directive, which was introduced to harmonise minimum Internet intermediary' liability requirements throughout the EU[22], including the general framework suitable for all

[16] Chow (n 14) 1299.
[17] Bayer (n 2) 5.
[18] Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market 2000.
[19] Sophie Stalla-Bourdillon, 'Internet Intermediaries as Responsible Actors? Why It Is Time to Rethink the E-Commerce Directive as Well' in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers*, vol 31 (Springer International Publishing 2017) 4 <http://link.springer.com/10.1007/978-3-319-47852-4_15> accessed 7 July 2020.
[20] Tambiama Madiega and others, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In-Depth Analysis.* (2020) 4 <https://op.europa.eu/publication/manifestation_identifier/PUB_QA0420239ENN> accessed 31 July 2020.
[21] Alexandre De Streel and others, 'Online Platforms' Moderation of Illegal Content Online' (2020) 11.
[22] Madiega and others (n 20) 1.

content and platform types. This general framework contains four different parts. Namely country of origin, immunity from liability for Online Platforms that are compliant and impartial and that remove illegal content online when it is identified, restrictions on broad surveillance measures to safeguard fundamental rights, promoting self-regulation and co-regulation and alternate conflict management structures.[23] The European Commission[24] has changed intermediary liability to expanded intermediary "responsibilities", believing that the position of online platforms is unique in its capacity to control the communication landscape and the associated relationship with users.[25] In order to further grow a web-based market around the EU and to encourage Online Platforms, the Commission also aims to reduce the legal uncertainty arising from the presence of conflicting network regulations between the different Member States.[26]

A revision has been made that determines and increases the obligations incumbent on Video Sharing Platforms (VSP) via the Audio-Visual Media Services Directive.[27] The dividing line between a VSP falling under the AVMSD and those which utilise limited liability for material under the E-Commerce Directive is becoming extremely blurred with so much audio-visual content now online.[28] According to this Directive, VSP was forced to adopt proportionality for racism, xenophobia and other hate speech, child sexual abuse, and terrorist content. The

[23] De Streel and others (n 21) 11.
[24] European Commission, 'Tackling Illegal Content Online. Towards an Enhanced Responsibility of Online Platforms.' (2017) Communication COM(2017) 555.
[25] Giancarlo Frosio and Martin Husovec, 'Accountability and Responsibility of Online Intermediaries' in Giancarlo Frosio (ed), Giancarlo Frosio and Martin Husovec, *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 1 <http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780198837138.001.0001/oxfordhb-9780198837138-e-31> accessed 3 August 2020.
[26] European Commission, 'Online Platforms and the Digital Single Market. Opportunities and Challenges for Europe' (n 4) 5.
[27] Directive 2010/13 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 audio-visual media services (Audio-Visual Media Services Directive), OJ [2010] L 95/1, as amended by Directive 2018/1808.
[28] Sally Broughton Micova, 'The Audiovisual Media Services Directive: Balancing Liberalisation and Protection' in Elda Brogi and Pier Luigi Parcu (ed), *Forthcoming Handbook on EU Media Law and Policy* (Edward Elgar Publishing 2020) 1 <https://ssrn.com/abstract=3586149>.

guarantee is determined according to the persons to be protected, and the feature and prevalence of the service(s) provided.[29] This directive is an example of European Union intermediary liability policy shifting from liability to responsibility.[30]

The Code of Conduct[31], signed by the most prominent online platforms, was initated by the Commission in 2016. The Code of Conduct focusses more on the prompt deletion of alleged hate language than on any procedural guarantees which should be adopted by such a private enforcement mechanism so as not to unexpectedly limit user freedom of speech. In the second part of this paper, the EU legal framework of online hate speech content will be examined and then the EU's legal framework problems will be analysed.

Currently, Online Platforms are gatekeepers[32] that "monitor what information we access and the conditions on which this material can be viewed," whereas users "lack the expertise and capacity to exert regulation" due to lack of awareness about how this monitor is, in fact, exerted.[33] Thus, their moderating ability makes them ethically responsible for the hate speech problem. They can be an essential part of solving/reducing the issue of hate speech. Today, the problem is not only whether social platforms can be planned and controlled, but also how the growing digital ecosystem can be conceptualised and controlled.[34] Many global and regional concepts of hate speech reflect on hate speech's various aspects. These various concepts hinder a clear and effective international regulatory system.[35] Online platforms are capable of both

[29] De Streel and others (n 21) 15.
[30] Aleksandra Kuczerawy, 'General Monitoring Obligations: A New Cornerstone of Internet Regulation in the EU?', *Rethinking IT and IP law: celebrating 30 years CiTiP* (2019) 1.
[31] 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (*European Commission - European Commission*) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 28 July 2020.
[32] Mostert (n 6) 609.
[33] Orla Lynskey, 'Regulating "Platform Power"' [2017] SSRN Electronic Journal 10 <http://www.ssrn.com/abstract=2921021> accessed 31 July 2020.
[34] Tiffany Li, 'Beyond Intermediary Liability: The Future of Information Platforms - Workshop Report' [2018] SSRN Electronic Journal 4 <https://www.ssrn.com/abstract=3392438> accessed 7 July 2020.
[35] Casarosa (n 10) 392.

monitoring the flow of information in their systems and removing the content they want.[36] There has been an increasing pattern in recent years for policymakers to utilise online platforms' terms of service to require that intermediaries delete illegal hate speech content.[37] The usage of takedown methods is also quicker and simpler for policymakers than legislative procedures to restrict or remove illegal content.[38] Content moderation may have shortcomings in the process of deleting or disabling links to illegal online content; such as being lengthy, untransparent etc. [39] The last chapter is primarily centred on online platforms aimed at developing legal structures to clarify their roles by corporate and social responsibility.

1. Background

1.1.Definition and Brief History of Hate Speech

The freedom of expression's boundaries are defined in Article 10 of the European Convention on Human Rights (ECHR) and Article 11 of The EU Charter[40]. It is considered one of the pillars of a democratic society and an essential precondition for ensuring other human rights are protected.[41] [42] Article 10 includes guarantees not only for harmless and non-intrusive speech but also against stinging and annoying speech. [43] Although freedom of expression is recognised as a fundamental human right, not all modes of expression are covered. In particular, types of speech that propagate, inspire, promote or excuse intolerance-based hate.[44] In other situations and in cases with hate content, the associated restrictions may be enforced.[45] The Committee

[36] Orit Finchman Afori, 'Online Rulers as Hybrid Bodies: The Case of Infringing Content Monitoring' (2020) 23 University of Pennsylvania Journal of Constitutional Law 1, 2.
[37] Li (n 34) 6.
[38] ibid.
[39] Rozenfeldova and Sokol (n 3) 867.
[40] Charter of Fundamental Rights of the European Union 2012.
[41] UNESCO, 'Freedom of Expression: A Fundamental Human Right Underpinning All Civil Liberties' (*UNESCO*, 14 April 2015) <https://en.unesco.org/70years/freedom_of_expression> accessed 29 July 2020.
[42] European Court of Human Rights, 'Factsheet – Hate Speech' (2020) 1 <https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf>.
[43] ibid.
[44] *Gündüz v Turkey* [2004] European Court of Human Rights Application no. 35071/97 76.
[45] Casarosa (n 10) 392.

of Ministers of the Council of Europe published a Declaration on Freedom of Communication on the Internet in 2003.[46] The Declaration makes a critical assessment of the competitiveness of the rights[47]. The Committee says that freedom of information, freedom of expression, and freedom of communication on the Internet should not be used in a way that would prejudice human dignity and human rights. [48] The increase in hate speech globally, which has been a significant challenge to democracy even in developed democracies triggered by the UN Secretary-General and caused the UN Strategy and Plan of Action on Hate Speech to be brought into existence.[49]

The European Court of Human Rights rulings echoed in several areas that barring discriminatory words in a democratic culture is essential and that hate expression is not safeguarded by Article 10 of the European Convention for the Protection of Human Rights and Fundamental Freedoms, which allows for freedom of speech.[50] The laws on harmful or illegal material differ due to their national, political, legal and religious backgrounds according to country,[51] although, it was declared at the 607th meeting of the Ministers' Deputies by the Council of Europe that a formal concept of hate speech would be used in the first and only foreign intergovernmental body.[52] The first article that we encounter that considers hate speech in the international community is Article 20 of the International Covenant on Civil and Political Rights[53]. The definition of this article is limited; indeed, so much so that it is related to national,

[46] Committee of Ministers, 'Declaration on Freedom of Communication on the Internet' (2003) 840th meeting of the Ministers' Deputies.
[47] Natalie Alkiviadou, 'Hate Speech on Social Media Networks: Towards a Regulatory Framework?' (2019) 28 Information & Communications Technology Law 19, 22.
[48] Committee of Ministers (n 46).
[49] 'United Nations Office on Genocide Prevention and the Responsibility to Protect' <https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml> accessed 30 July 2020.
[50] European Court of Human Rights (n 42).
[51] Yu Wenguang, 'Internet Intermediaries' Liability for Online Illegal Hate Speech' (2018) 13 Frontiers of Law in China 342, 345.
[52] Bard and Bayer (n 12) 30.
[53] 'OHCHR | International Covenant on Civil and Political Rights' <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx> accessed 30 July 2020.

racial and religious descriptions. Hate speech requires the advocacy of international hatred that provokes people to violence or discrimination. With this agreement, states have been obliged to provide an effective remedy to its victims, regardless of who commits hate speech, whether state agent or citizen.[54] The issue of online hate is not addressed directly in Article 4 of the International Convention on the Elimination of All Forms of Discrimination. Still, this convention outlaws specific forms of expression, and racial discrimination instead.[55] The two Conventions, as mentioned above, were written before the use of the web became widespread, [56] which is why they did not otherwise consider online hate speech.[57] In 2003, another indication of the universal agreement that different kinds of online hate speech would be banned by regulation was the "Protocol on the Criminalisation of Acts of a Racist or Xenophobic Nature" as provided by the Council of Europe. However, this is not enough to prevent online hate since it addresses solely racist and xenophobic speech.

Although the Framework Convention for the Protection of National Minorities[58] does not contain specific wording of hate speech, according to its 6th article. Parties promoting intercultural communication should take reasonable steps to defend persons who, due to their national, racial or religious status, are vulnerable to danger or bigotry, aggression or abuse. The Additional Protocol to the Convention on Cybercrime[59] purposes harmonising the criminalisation of actions of racism and xenophobia performed via computer devices. However, it does not consider anything but issues of racism and xenophobia.[60] Paragraph 25

[54] Bard and Bayer (n 12) 28.
[55] 'OHCHR | International Convention on the Elimination of All Forms of Racial Discrimination' <https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx> accessed 7 August 2020.
[56] Alkiviadou (n 47) 27.
[57] ibid.
[58] 'Framework Convention for the Protection of National Minorities - European Treaty Series - No:157' (*CoE Treaty Office*) <https://www.coe.int/en/web/conventions/full-list> accessed 30 July 2020.
[59] 'The Additional Protocol to the Convention on Cybercrime Treaty No.189' (*CoE Treaty Office*) <https://www.coe.int/en/web/conventions/full-list> accessed 30 July 2020.
[60] Alkiviadou (n 47) 23.

of this protocol ensures liability immunity for service providers if their interference is not sufficient to merit condition.

The Committee of Ministers Recommendation on Hate Speech's[61] Principle 3 points out that "any restriction or conflict with freedom of speech must be subject to impartial judicial oversight" which is especially important in light of social media legislation. Because hate speech may reflect serious forms of discrimination, ECHR's Article 12 and Protocol 12, describe the prohibition of discrimination, which is especially relevant with regard to the issue of hate speech. Enforcing liability for illegal hate speech on online platforms requires a precise description of hate speech in the applicable legislation and platforms.[62] And, as can be seen in this chapter, there is no precise description. Such conclusive confusion may encourage online platforms to over-block[63] and suppress expression.[64]

Although the Counter-Racism Framework Decision aims to counter especially troubling manifestations of xenophobia and racism by criminal legislation, it does not describe xenophobia and racism or include xenophobic and racist hate speech.[65] Explicitly, advocating, ignoring or underestimating war crimes, genocide and crimes against humanity are criminalised by the Counter-Racism Framework Decision if aimed at a category of individuals or a member of such a community as described in relation to national origin, descent, colour, ethnicity or religion. However, Member States are free to regulate a broader set of points so covered in addition to those described in the Counter-Racism Framework Decision [66] such as sexual orientation, language, belief and political opinion. Arousing hatred or brutality are also

[61] Committee of Ministers, 'Recommendation No. R (97) 20 of The Committee Of Ministers To Member States on "Hate Speech' (1997).
[62] Wenguang (n 51) 353.
[63] Mostert (n 6) 612.
[64] Wenguang (n 51) 353.
[65] De Streel and others (n 21) 19.
[66] Bard and Bayer (n 12) 50.

criminalised by Counter-Racism Framework Decision too.[67] Although this paper is undoubtedly a significant move forward in overcoming discrimination and hatred, it is often criticised for being excessively conservative in terms of freedom of speech.[68]

Although Online platforms usually complain about disproportionately and different regulations,[69] online platforms define hate speech very broadly. Online platforms' strategies in this regard will be examined in chapter 3.1.

1.2. Governments' Approaches to Liability of Social Media Companies for Hosting Prohibited Content on their Platforms

The liability problem regarding content from third parties had already began by the 1990s.[70] The internet service provider denied the alleging defamation in the *Godfrey v Demon Internet Ltd*[71] case, in which Godfrey believed that an anonymous user-produced posting that contained defamatory and obscene material, had fraudulently credited him as an author. This posting was published on an online platform operated by Demon Internet Limited, who did not delete the post instead allowed it to remain available for 20 days until its expiry date.[72] Godfrey's application to the courts succeeded.

1.2.1. Extended Liability Immunity

The first alternative that policymakers have is not to hold Online Platforms liable for hate speech on their websites. Immunity from intermediary liability was initially adopted for the

[67] De Streel and others (n 21) 19.
[68] Bard and Bayer (n 12) 51.
[69] Li (n 34) 9.
[70] Bayer (n 2) 12.
[71] *Laurence Godfrey v Demon Internet Limited* [2001] QB [1999] EWHC QB 244.
[72] 'Godfrey v. Demon Internet Limited' (*Global Freedom of Expression*) <https://globalfreedomofexpression.columbia.edu/cases/godfrey-v-demon-internet-limited/> accessed 31 July 2020.

aim of supporting and innovating an evolving internet industry.[73] Such regulatory provision leaves online players with broad flexibility on the matter. It does not, in essence, offer any explicit instructions.[74] Online platforms are shielded from responsibility for material shared by users since they are not regarded as publishers.[75] For instance, the U.S. supports Online Platforms to remove hate posts without holding them liable.[76]

One of the reasons for defending this view is that if the liability risk increases, small businesses will be financially unable to employ the number of moderators or lawyers required to cope with it, so this risk will eliminate the competition, and start-ups will not be supported under this financial burden.[77] New requirements placed on online platforms raise competition obstacles to access and survive for the market.[78] The other benefit of this approach is that of respecting users' right to freedom of speech. The relationship between freedom of expression and liability will be examined in detail the following chapters.

The U.S. allows extended immunity from liability for content by third parties through section 230 of "the Communications Decency Act 1996 which provides intermediaries with extended security against liability for the content created by users so long as they do not interfere with the content, acting as "good Samaritan" in good faith to block and filter offensive material.[79] With regard to this strategy, intermediaries were able to establish market strategies focussed on freedom of knowledge dissemination.[80] Online platforms offer voluntary monitoring

[73] Giancarlo F Frosio, 'Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy' (2017) 4 Centre for International Intellectual Property Studies Research Paper 7.
[74] Afori (n 36) 13.
[75] Communications Decency Act of 1996, 47 U.S.C. § 230(c) (2012)
[76] Chow (n 14) 1301.
[77] Daphne Keller, 'Toward a Clearer Conversation About Platform Liability' (Knight First Amendment Institute's 2018) 1.
[78] Christophe Geiger, Giancarlo Frosio and Elena Izyumenko, 'Intermediary Liability and Fundamental Rights' (Center for International Intellectual Property Studies 2019) 6 19.
[79] Keller (n 77) 2.
[80] Andrej Savin, 'New Directions in EU Policymaking on the Content Layer: Disruption and Law' (Copenhagen Business School Law 2020) 20–05 5.

activities to provide a "civilised" environment on their platforms as an indicator of being Good Samaritan.[81]

The United States Court of Appeals for the First Circuit ruled that a website containing pornography ads involving minors coerced into trafficking was not liable because it was not deemed a "publisher" of such third-party material.[82] The U.S. District Court for the Eastern District of New York ruled that although Facebook was used to promote, recruit, mobilise and submit individuals to attack Israelis by the Palestinian terrorists, Facebook is still not liable,[83] because Facebook was still not the "publisher" of that content.[84] With this strategy, hate speech will stay available for longer. Besides, it is said that companies earn advertising revenue from user content, thus they must also cover the costs of their activities.[85]

Although Social Platforms have broad immunity under Section 230, they have established filtering systems and monitoring mechanisms to voluntarily and actively control,[86] because it was economically lucrative for them to create an environment that reflects the expectations and norms of its users, even though this is not required by the law.[87]

1.2.2. Holding Completely Liable

In this model of liability, online platforms are held liable for nearly all activities of third parties, as online platforms' structures facilitate the dissemination of hate speech and other offensive content, and allow more practical grounds of potential aggression[88] against different

[81] Kuczerawy (n 30) 2.
[82] *Jane Doe No 1 v BackpageCom, LLC* [2016] United States Court of Appeals For the First Circuit 817 F.3d 12 (1st Cir. 2016).
[83] *Cohen v Facebook, Inc* [2017] United States District Court, ED New York 252 F. Supp. 3d 140 (E.D.N.Y. 2017).
[84] Chow (n 14) 1301.
[85] *Delfi As v Estonia* [2015] The European Court of Human Rights Application no. 64569/09.
[86] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2017) 6 Harvard Law Review 131, 161.
[87] ibid 150.
[88] Stuart (n 9) 127.

communities.[89] Chinese intermediary liability system is the biggest example of this method.[90] China's Internet Regulation Strategy and intermediary liability system give an aggressive surveillance responsibility to Intermediaries.[91]

This strategy has been criticised by scholars because it requires social networking sites to voluntarily(!) delete illegal content, which would mean freedom of expression could easily be adversely affected.[92] Online platforms may find themselves having a decision about whether to protect themselves and delete content which may not be accepted as illegal content.

1.2.3. Liability Under Specific Conditions

This method is more suitable for the role of Online Platforms in that it is facilitation, not dissemination.[93] In other words, the third approach accepts that Online Platforms are just intermediary functions, rather than editing abilities. As will be examined below, the E-Commerce Directive is a famous example of this approach. In this liability system, there is a safe harbour system, whereby online platform users can not be subject to general monitoring by online platforms, because this might be harmful to the fundamental rights of users.[94]

2. EU Regulatory Framework on Online Hate Speech Moderation

The 3 basic features of EU content regulation can be considered. Laws have a proper legal basis, proportionality and subsidiarity.[95] While reaching the Digital Single Market, which is

[89] Chow (n 14) 1303.
[90] Wenguang (n 51) 347.
[91] ibid 348.
[92] Chow (n 14) 1303.
[93] ibid 1306.
[94] Madiega and others (n 20) 1.
[95] Savin (n 80) 13.

the final goal of the EU, it was trying to avoid unnecessary laws, and it was aiming to make regulations taking into account the nature of the regulated issue. [96]

2.1.The E-Commerce Directive

2.1.1. Hosting Provider's Privileges and Obligations on ECD

Together with the E-commerce Directive (ECD), the online intermediary services entered a specific liability regime in 2000. The main targets of the ECD are that while keeping the rights of various actors such as privacy and freedom of speech, the freedom to conduct the business of online platforms would be given a fair balance, to provide Union harmonisation with regard to the digital single market, to revive the e-commerce within the Union by not holding online platforms liable for monitoring the legality of all the content they might host, a cooperation of public officials and private actors for a secure internet.[97] The liability exemption of the ECD embraces both criminal and legal proceedings. It includes member states' liability regimes to which the platforms are subject.[98]

Article 12 of the ECD is related to liability of 'mere conduit service providers'. In this article, service providers have allow transmission activity in a passive manner. Caching providers are offered immunity from Article 13 if they store online materials automatically and temporarily to transmission more effectively. In article 14, hosting services are exempt from liability if they do not realise they are hosting a user's illicit activity or content and respond expeditiously to remove or prevent access to illegal content, such as, take down and block. In this way, injured parties inform the online platforms of any illegal situation,[99] so that they can request that the

[96] ibid.
[97] Commission, 'Explanatory Memorandum of the Commission proposal for a directive on certain legal aspects of electronic commerce in the internal market', COM(1998)586.
[98] De Streel and others (n 21) 23.
[99] Sophie Stalla-Bourdillon and Robert Thorburn, 'The Scandal of Intermediary: Acknowledging the Both/and Dispensation for Regulating Hybrid Actors' in B.Petkova and T.Ojanen (ed), *Fundamental Rights Protection Online: the Future Regulation of Intermediaries* (Edward Elgar Publishing 2019) 152.

offending content be removed or blocked. [100] However, here is neither a description of actual knowledge nor a description of illegal activity.[101] Their meanings depend on Member States' views. Thus, it will create such ambiguity. In order to avoid or discourage potential illegal conduct, Member States have the ability to develop or utilise established provisional relief against intermediary service providers in Articles 12, 13 and 14.[102]

If the platform acts expeditiously to delete or disable information access, it will still benefit from the exemption from liability even if it has knowledge or details. However, when the platform receives such a notification, how can it balance the rights of the content owner and the person claiming that their rights have been harmed? This issue is not clear in the Directive.[103] Besides, Member States can determine a notice, and takedown/stay down/notice method according to their preferences. This may create uncertainty for Online Platforms. Article 15 of the ECD bars the enforcement of a general obligation on the part of the hosting platforms to monitor the hosted content for the EU Member States. The requirement to comply with the relevant authorities is placed on hosting services by the Member States. The discretion and transparency of platform-based private regulatory systems is a significant concern.[104] Thus, nothing prohibits Internet intermediaries from actively participating in general monitoring, even though general monitoring is banned.[105] In order to significantly restrict the legislative power of social networking companies in general monitoring, Article 15 must be implemented through audits. According to Article 28(b) AVMSD, stricter measures to deal with illegal

[100] Commission, 'Explanatory Memorandum of the Commission proposal for a directive on certain legal aspects of electronic commerce in the internal market', COM(1998)586.
[101] Madiega and others (n 20) 5.
[102] Stalla-Bourdillon and Thorburn (n 99) 153.
[103] De Streel and others (n 21) 23.
[104] Stalla-Bourdillon and Thorburn (n 99) 157.
[105] Giancarlo Frosio, 'The Death of "No Monitoring Obligations": A Story of Untameable Monsters' (2017) 8 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 199, 200.

content may be required by the Member States from VSP's[106]. However, what might constitute more restrictive behaviour is ambiguous in AVMSD. Thus these measures would be inconsistent with Article 15 of the guideline on e-commerce.[107]

2.1.2. European Courts' Approach to the Liability Exemption

The Court of Justice of the European Union made a critical assessment of the extent to which the service provider could benefit from the exemption from *the Google France v Louis Vuitton*[108] and *the L'Oreal et al. v. eBay*[109] decisions. According to these decisions, if the service provider is in a neutral role, the service provider shall not be responsible for the content retained at the discretion of the user.[110] If the service provider has the power to analyse the data given to them by clients and decide in what sequence advertisements should be shown, it will therefore no longer be excluded from liability.[111] For instance, to optimize or encourage the appearance of the related selling deals by drafting the promotional post or consciously choosing the corresponding keywords.[112]

It is stated in the *Google France v. Louis Vuitton* decision that the service provider can be considered neutral if it plays only in a passive, technical and automated role. However, in the *L'Oreal et al. v. eBay* decision that it is not possible for a service provider to benefit from any exemption to liability, since it plays an active role between the customer and the seller. The Court's decisions regarding the exception of liability, which is required to be of a passive and

[106] 'Regulating Content Moderation in Europe beyond the AVMSD' (*Media@LSE*, 25 February 2020) <https://blogs.lse.ac.uk/medialse/2020/02/25/regulating-content-moderation-in-europe-beyond-the-avmsd/> accessed 9 August 2020.
[107] Stalla-Bourdillon and Thorburn (n 99) 163.
[108] C-236/08 to C-238/08 Google France v Louis Vuitton EU:C:2010:159.
[109] Case C-324/09 L'Oreal et al. v. eBay EU:C:2011:474.
[110] De Streel and others (n 21) 22.
[111] Rozenfeldova and Sokol (n 3) 872.
[112] Lisl Brunner, 'The Liability of an Online Intermediary for Third Party Content The Watchdog Becomes the Monitor: Intermediary Liability After' (2016) 16 Human Rights Law Review 163, 7.

automated nature[113], creates uncertainty about proactive monitoring of materials legality hosted by online platforms[114]. It should be remembered that a case-by-case analysis is most frequently required in order to decide if a certain provider's activities should be considered services.[115]

A news article published in 2006 by Delfi, the Estonian news website, about a shipping company called SLK, amassed 20 comments containing threats and insults about L, who is a member of the SLK's board of directors. Although Delfi removed the racial hate speech content of third-parties after notification, it was disciplined by the national courts,[116] because Estonian national courts state that the Directive's Articles 12-14 did not refer to Delfi.[117] Based on the premise that the plaintiffs are responsible for content by third parties, the European Court of Human Rights rendered its decision[118] since Delfi has the ability to exercise a significant degree of discretion over the remarks reported on its site.[119] The ECHR also found that Delfi's liability for the content of third parties did not violate its freedom of expression under Article 10.[120] According to ECHR's Delfi As ruling, Internet intermediaries must erase individuals' defamatory remarks, and they will be found responsible for any such derogatory statements. Judges Sajo and Ttostoria state that intermediaries may have substantial reasons to avoid allowing users opinions due to fear of liability. This will open the gates to further self-censorship.[121] The judgement of the European Court in respect of unlawful material produced by users without awareness of the assumption of liability on an on-line news source is difficult to comply with ECD's liability approach.[122] An accurate decision might be that Delfi conducts

[113] Stalla-Bourdillon and Thorburn (n 99) 152.
[114] De Streel and others (n 21) 22.
[115] Rozenfeldova and Sokol (n 3) 871.
[116] *Delfi As v. Estonia* (n 85).
[117] Bayer (n 2) 16.
[118] *Delfi As v. Estonia* (n 85).
[119] ibid 130–145.
[120] Daithí Mac Síthigh, 'The Road to Responsibilities: New Attitudes towards Internet Intermediaries' (2020) 29 Information & Communications Technology Law 1, 13.
[121] *Delfi As v. Estonia* (n 92) dissenting opinion para. 1.
[122] Brunner (n 112) 6.

its duties by rigorous processes of due diligence.[123] Such as deleting defamatory comments when it becomes aware, using notice and takedowns systems.

2.2.2018 Recommendations of the Commission

A Recommendation[124] for hosting platforms and the Member States was adopted by the European Commission in March 2018 to take efficient, proactive, appropriate and proportionate measures to counter illegal content online. It describes the principles of all sorts of illegal online contents and advises that terrorist content be more rigidly moderated.

Recommendations can be grouped under three categories. 'notice-and-takedown', proactive measures and cooperation. According to the Recommendation's points 5 to 17, the 'notice-and-takedown' method should be transparent, effective, adequately accurate and proved, should consider the rights of content creators and offer opportunity for 'counter-notices' and out-of-court dispute resolutions. According to the Recommendation's points 16 to 21, ‘proactive measures’ should contain the usage of digital instruments, through human monitoring and assurance as a safeguard, for appropriate, proportionate, and precise interventions. According to the Recommendation's points 22 to 28, tight cooperation should occur with trusted flaggers, especially for smaller service providers who might be less capable of coping with illegal content, and national judicial and administrative authorities on a transparent and fair basis.

The period when these measures are implemented is essential to remember because actions performed solely before the incident happens can be accepted as proactive measures.[125]

2.3.The Audio-Visual Media Services Directive

[123] ibid 8.
[124] European Commission, ‘Recommendation 2018/334 on Measures to Effectively Tackle Illegal Content Online’ (2018) OJ [2018] L 63/50.
[125] Rozenfeldova and Sokol (n 3) 876.

In November 2018, the AVMSD was introduced, which was aimed at properly understanding the online world and providing more equitable opportunities between traditional media channels and emerging video sharing platforms. From Article 28(b) of AVMSD, there is a clear effort to create a framework that is in compliance with Articles 12 to 15 of the ECD. AVMSD's extended scope involves matching threats to freedom of speech with crucial customer protection and the goal of preserving national media enterprises, which are essential in Member States' cultural and political being.[126] According to Article 28 of AVMSD, effective restrictions on Video-Sharing Platforms with the purpose of protecting society from hate speech, child sexual abuse material, terrorist content and racism and xenophobia that are forbidden under EU legislation and to shield children from material that may harm their physical, emotional or moral growth are required.[127]

Furthermore, AVMSD describes the potential steps shall be adopted by VSP's such as open and user-friendly monitoring and flagging systems[128] enabling content rated by the user, and efficient processes for solving users' complaints.[129] The AVMSD marks a significant shift because "real, enforceable rights" would be provided for the first time for a VSP's users.[130] Besides, in consideration of its existence, its possible injury, the status of individuals to be safeguarded, the freedoms, and the valid interests at risk, the AVMSD stipulates that the measures should be appropriate.[131] In relation to the scale of the VSP and the essence of the service delivered, the provisions need to proportionate.[132] Therefore, the controls levied on VSPs cannot result in *ex-ante* monitoring steps or content upload filtering.

[126] Sally Broughton Micova (n 28) 1.
[127] Mac Síthigh (n 120) 7.
[128] Sally Broughton Micova (n 28) 16.
[129] Lubos Kuklis, 'Video-Sharing Platforms In AVMSD – A New Kind Of Content Regulation', *Research Handbook on EU Media Law and Policy* (Elgar Publishing 2020) 18.
[130] 'Regulating Content Moderation in Europe beyond the AVMSD' (n 106).
[131] De Streel and others (n 21) 25.
[132] Kuklis (n 129) 19.

The AVMSD has been criticised by some scholars that it does not maintain the obligations or other general liabilities established toward users of VSPs.[133] According to Article 28b, member states may require the VSP to apply more stringent and detailed measures than those mentioned above. It is not clear whether the more stringent and detailed measures here mean the *ex-ante* filtering prohibited by Article 15 of the ECD.[134] The steps suggested include striking a fair balance between the desire to strengthen the freedom to conduct business on the one hand and the need to uphold human liberties on the other.[135] The AVMSD reflects a possible restriction for immunity to liability of intermediaries, and changes the theme from one of liability to one of responsibility.[136] The application of laws leading to the exclusion of some kinds of materials which are not actually unlawful appears to be encouraged by 'integration' into the conditions of the customer within the providers.[137]

2.4.The Code of Conduct

The EU Code of Conduct (CC) to counter unlawful hate speech online was accepted by Snapchat, YouTube, Microsoft, Jeuxvideo.com, Instagram, Twitter, Facebook, Google+ and Dailymotion at the leadership of the European Commission. (in 2016 for the first OP and in 2019 the last five OPs).[138]

The CC finds that online platforms play a vital role in maintaining enforcement, and a number of responsibilities have been entered into by the platforms.[139] Firstly, attracting focus from users to content forms which are not approved by their Community Standards /Guidelines and

[133] ibid 8.
[134] Maria Lillà Montagnani, 'A New Liability Regime for Illegal Content in the Digital Single Market Strategy' (Bocconi University 2019) 477007 16.
[135] De Streel and others (n 21) 25.
[136] Frosio (n 73) 1.
[137] Stalla-Bourdillon and Thorburn (n 99) 163.
[138] 'The EU Code of Conduct on Countering Illegal Hate Speech Online' (n 31).
[139] De Streel and others (n 21) 30.

barring incitation to aggression and offensive behaviour. Secondly, simple and appropriate mechanisms are developed for the analysis of incidents/alerts of unlawful or unavailable hate speech; alerts are examined in accordance with Community standards/guidelines and national transposition law. Thirdly, digital portal workers are routinely taught, in particular about social developments. Fourthly, enhance coordination and cooperation between NGOs, OP's and governments, share best practices and finally facilitate reporting by experts on hate speech. Although civil society has not been involved in the drafting process sufficiently, NGOs have been active in checking the framework such as flagging illegal material and providing the European Commission with details about the conformity of their applications.[140]

The compliance and effect of the CC have periodically been measured by the Commission on the basis of details received by the platforms. The fifth evaluation of June 2020 shows that 90.4 per cent of notifications are reviewed within 24 hours. The goal of updating alerts within one day is entirely achieved by both information technology (IT) organisations, and the pattern of success relative to the previous reporting exercise persists.[141] IT firms have also withdrawn 71 per cent of the material disclosed to them.[142] Facebook deleted 87.6 per cent of content, YouTube 79.7 per cent, and Twitter 35.9 per cent.[143] The evaluation reports show that Ops responded to the flagging more quickly last year and deleted a larger amount of flagged material than in previous years. Nevertheless, no systematic research has been undertaken to date about what is behind the numbers. The findings will not disclose the specifics of the OP's decisions.[144] This may cause trouble for researchers and NGO's in terms of studying the removal mechanism.

[140] Bard and Bayer (n 12) 53.
[141] Didier Reynders, '5th Evaluation of the Code of Conduct' 5, 2.
[142] ibid.
[143] ibid.
[144] Bard and Bayer (n 12) 54.

One of the most significant criticisms related to the CC is that authorities might force private corporations to remove material which is not allowed by these authorities to remove it.[145] The following shortcomings about the CC were pointed out: failure to assess the authenticity of notice, lack of the frameworks of appeal for users of removed content, once deleted on the grounds of Community Standards/Guidelines illegal material will not have to be reported to the appropriate national authorities, this 24-hour span may either prohibit online platforms from meeting their obligations or drive them to over-blocking activities. [146] It is pointed out that the efficacy measure depends on the amount and level at which the items withdrawn are eventually erased, not their illegality.[147] It is assumed that the CC may encourage individuals to switch to another OP that is less controlled because the EU opt for censorship rather than tackling the underlying causes of hate speech and the societal issues concerned.[148]

3. Online Platforms' Policies and Tools to address Hate Speech Content

Intermediaries have the capacity to delete and monitor illegal hate speech material with assistance of technological developments and financial resources.[149] How can OPs deal with hate speech? They can use terms of services, filtering, manipulation of searches, etc.

3.1.Terms of Service

Online platforms provide services to their users by preparing the Terms of Service. These conditions generally consist of elements such as who will be the owner of the contents' copyrights created by the users, privacy issues, the responsibility of the content and the

[145] ibid.
[146] Teresa Quintel and Carsten Ullrich, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond' in BILYANA PETKOVA and TUOMAS OJANEN (eds), *Fundamental Rights Protection Online: the Future Regulation of Intermediaries* (Edward Elgar Publishing 2019) 16.
[147] ibid.
[148] Barbora Bukovská, 'The European Commission's Code of Conduct for Countering Illegal Hate Speech Online' [2019] Transatlantic Working Group, 7.
[149] Wenguang (n 51) 350.

responsibility arising from the content, the choice of jurisdiction and law.[150] Online platforms produce Terms of Service to control and monitor users' behaviour and to form their foundation to use enforcement against the user who creates illegal content.[151] This system may also have methods of complaint for reporting illegal content. The Terms of Service are criticised for being more rigid than stipulated in the detection of illicit materials that are to be deleted electronically than national laws. [152] Such terms of services' go beyond the requirements of national legislation and also universal norms in general (in Europe, at least).[153] Yet, UN reporter says to governments and companies that to classify the type of material that identifies as hate speech with rational consumer and public examples and coherent strategies across jurisdictions.[154]

Facebook defines hate speech as a "direct attack" on people based on what we call protected characteristics - race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disease or disability."[155] Google adds immigration status and veteran status to the above. [156] YouTube has a broad scope for hate speech. YouTube prohibits dehumanising, conspiracy theories attacking particular groups, etc.[157] This may be interpreted as the achievement of a higher level of protection.[158] Twitter alerts its users about hate speech, rather than forbids it.[159] In addition, Twitter does not mention hate speech in it

[150] Corinne Tan, *Regulating Content on Social Media: Copyright, Terms of Service and Technological Features* (UCL Press 2018) 99 <http://www.jstor.org/stable/10.2307/j.ctt2250v4k> accessed 14 August 2020.
[151] De Streel and others (n 21) 40.
[152] ibid 43.
[153] 'Regulating Content Moderation in Europe beyond the AVMSD' (n 106).
[154] 'OHCHR | Governments and Internet Companies Fail to Meet Challenges of Online Hate – UN Expert' <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25174&LangID=E> accessed 9 August 2020.
[155] Facebook, 'Community Standards' <https://www.facebook.com/communitystandards/hate_speech> accessed 29 July 2020.
[156] 'Hate Speech Policy - YouTube Help' <https://support.google.com/youtube/answer/2801939?hl=en> accessed 29 July 2020.
[157] ibid.
[158] Casarosa (n 10) 398.
[159] Alkiviadou (n 47) 24.

terms of service, merely referring to "offensive, harmful, inaccurate or otherwise inappropriate..."[160]

Even though states do not hold companies liable from third party content, companies might aim to overcome hate speech and harmful content because it is unlikely that a website filled with hate speech will attract users. For instance, the YouTube Partner Program started by YouTube in 2017 motivates users to monetise appropriate content with ads. [161] In order to this program, If users' content is not accepted suitable for Advertiser-friendly content guidelines, user's can not benefit from ads.[162]

3.2.Searching Manipulation

Search manipulation deletes illegal online contents from search results. For instance, Google introduced a system to hide revenge porn related actions.[163] Facebook also created a tool to protect people who are victim of revenge porn. When Facebook decides that images are related to revenge porn, they immediately use picture matching technology to prevent further sharing on Messenger, Facebook and Instagram.[164]

3.3.Filtering and Content Recognition Tools

Filtering ensures that certain words or slogans are prevented from being published in the system and deleted if they are. The automatic control of the content prevents the rapid inspection of systems that produce and contain millions of users, vast amounts of content, and the spread of

[160] 'Twitter Terms of Service' <https://twitter.com/tos?lang=en#usContent> accessed 7 August 2020.
[161] 'Advertiser-Friendly Content Guidelines - YouTube Help' <https://support.google.com/youtube/answer/6162278> accessed 14 August 2020.
[162] ibid.
[163] 'Google to Exclude "revenge Porn" from Internet Searches | Google | The Guardian' <https://www.theguardian.com/technology/2015/jun/20/google-excludes-revenge-porn-internet-searches> accessed 3 August 2020.
[164] 'Using Technology to Protect Intimate Images and Help Build a Safe Community' (*About Facebook*, 5 April 2017) <https://about.fb.com/news/2017/04/using-technology-to-protect-intimate-images-and-help-build-a-safe-community/> accessed 3 August 2020.

hate speech.[165] This regulation tool with *ex ante* base may result in unpredictable censorship.[166] According to Article 46 of the E-Commerce Directive, when a takedown request is considered by the online platform, the platform must respect the principle of freedom of expression.

Google's content iD utilises digital fingerprints, and hashes, to match an uploaded file to a database of protected works provided by rights holders.[167] There has been no significant progress with other methods to identify offensive and provocative material. OPs use filters, including certain words, sentences or hashes are sometimes too uninclusive to detect unlawful material.[168] Creating new wordings is easy, and meaning is more essential than wordings. Daily language mainly contains humour, sarcasm and irony.[169] 99,5 per cent of extremist material removal, 96 per cent of pornography and pornographic material removal, and 86% of abusive content removal was already immediately effected by Facebook in 2018, but just 38 percent of the hate speech was removed in the same period.[170]

When an algorithm operates solely by the detection of such keywords, the sophistication of human speech can not be mastered because, with the lack of meaning, it is likely to generate unintended false positives and negatives.[171] False positives are also a crucial issue for this method. Human moderators may overcome false positive problems.[172] Another problem with filtering hate speech is that some words are 'hyperlocal'.[173] Some words, for example, may only have meaning for people in a particular town or region. Therefore, online platforms should

[165] Jennifer Cobbe, 'Algorithmic Censorship by Social Platforms: Power and Resistance' 3 <https://ssrn.com/abstract=3437304>.
[166] Geiger, Frosio and Izyumenko (n 78) 13.
[167] 'How Content ID Works - YouTube Help' <https://support.google.com/youtube/answer/2797370?hl=en> accessed 3 August 2020.
[168] Chow (n 14) 1304.
[169] Cobbe (n 165) 3.
[170] Jason Koebler and Joseph Cox, 'Here's How Facebook Is Trying to Moderate Its Two Billion Users' <https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works> accessed 12 August 2020.
[171] Casarosa (n 10) 396.
[172] De Streel and others (n 21) 44.
[173] Arun (n 11) 2.

be in dialogue with local authorities. Neither AI nor human moderators who live in another country can correctly evaluate such phrases and words without local support.[174]

The CJEU[175] ordered that an online platform, such as Facebook, could be required to detect and remove illicit defamatory statement and 'equivalent statements' made by the same person. With this decision, online platforms can automatically censor content considered the equivalents of illegal contents.[176] This may results in general monitoring which may infringe freedom of speech.[177] As can be seen regarding this decision, the premise that a social network function can be deemed passive in nature is becoming extremely challenging to pursue.[178] If we bear in mind that general monitoring is prohibited by Article 15 of the ECD, uncertainty thus arises. However, according to the Advocate General[179], if a particular infringement has already been detected regarding content from a user, the service provider can monitor the same content from that particular user, and there is thus no generalized, public tracking. In addition, even in the absence of general monitoring, the service provider using excessive control could jeopardise its neutral status and, as a result, lose its safe harbour protection under Article 14 of the ECD.[180]

3.4.Three-Strike Mechanisms, Blocking Websites and Blocking Money Transfers

Online Platforms can not use these measurements, but other intermediaries (and states) can. Three-strike mechanisms aim to prohibit infringers who repeatedly act from connections of the household Internet.[181] This mechanism is applied by Internet service providers, and is intented

[174] ibid 4.
[175] *Eva Glawischnig-Piesczek v Facebook Ireland* [2018].
[176] 'The Glawischnig-Piesczek v Facebook Case: Knock, Knock. Who's There? Automated Filters Online' (*CITIP blog*, 12 November 2019) <https://www.law.kuleuven.be/citip/blog/the-glawischnig-piesczek-v-facebook-case-knock-knock-whos-there-automated-filters-online/> accessed 5 August 2020.
[177] ibid.
[178] Stalla-Bourdillon and Thorburn (n 99) 157.
[179] *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] paragraph 58.
[180] Kuczerawy (n 30) 2.
[181] Frosio and Husovec (n 25) 3.

to solve copyright infringements.[182] The State is liable under Article 10 ECHR and article 11 EU Charter to growing action that is expected to have an effect on the functionality of the internet.[183] If this functionality restriction would result in a Web site blocking, it should be borne in mind that if this is not done properly, it will be against freedom of expression. [184] Therefore, the institution that will decide on this restriction should be the courts. [185] And while these courts are making their decision, they need to look at the use of the website to be blocked and the effect of the block on communication. [186]

The freedom of speech can be subject to interference from payment intermediaries. [187] For instance, with regard to WikiLeaks, who have not faced any court proceedings to date, numerous payment methods, Visa, MasterCard, PayPal, etc., have suspended the transfer of donations to them.[188]

3.5.Notice and Takedown and Flagging

Various online platforms have suggested that this device is being implemented as the bulk of illicit material can not be identified online through electronic methods and can not be identified on a human scale.[189] Common account holders are not usually able to classify illegal web content correctly.[190] There are certain problems with finding complaint mechanisms. YouTube and Twitter allow the usual "flagging" interface available from direct links next to each piece

[182] 'New Education Programme Launched to Combat Online Piracy' (*GOV.UK*) <https://www.gov.uk/government/news/new-education-programme-launched-to-combat-online-piracy> accessed 3 August 2020.
[183] Geiger, Frosio and Izyumenko (n 78) 8.
[184] ibid.
[185] ibid.
[186] ibid.
[187] Frosio and Husovec (n 25) 6.
[188] 'Visa, MasterCard Move To Choke WikiLeaks' <https://www.forbes.com/sites/andygreenberg/2010/12/07/visa-mastercard-move-to-choke-wikileaks/> accessed 3 August 2020.
[189] De Streel and others (n 21) 44.
[190] ibid 51.

of content about which a complaint has been raised.[191] Youtube also has its own flagging systems to collect potentially appropate materials.[192] The complaint form on Facebook is on a different, less clearly illustrated page that requires several clicks to reach.[193] Facebook offers a rulebook to workers from a number of locations around the globe to help them understand community standards and make formal judgments on notice and takedown.[194]

Big online platforms, for instance, Facebook, Apple, Google and Amazon, even have the ability to engage in political bargaining to create diplomatic negotiating arrangements with governments. They may secure advantageous privileges from intermediary liability requirements where certain small firms can consider it challenging to obey.[195]

The result of a conservative approach to internet control is that banned material has stayed online for at least some period of time.[196] Facebook deleted a video displaying the assassination of Robert Godwin Sr. in Ohio two hours after it had been released. Facebook did not catch the prior video in which the killer described his plan, and nor did any users raise it as an issue.[197] In addition, concerning the increase in numbers of OP's users day by day, moderating all content is also an incredible burden for OP's attemting to protect themselves from possible prosecution. Thus, OP's have to ensure serious moderation with regard to deleting, even illegal content which they may be not aware of. OP's may use human moderators to deal with this problem. Human moderators often experience stress whose effects range from weakening their judgement and growing attrition, having to see thousands of horrific or vitriolic images each

[191] Heidi Tworek and Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' (Transatlantic High Level Working Group 2019) 5.
[192] Michael Patty, 'Social Media and Censorship: Rethinking State Action Once Again' (2019) 40 Mitchell Hamline Law Journal of Public Policy and Practice 99, 106.
[193] Tworek and Leerssen (n 191) 5.
[194] Patty (n 192) 106.
[195] Li (n 34) 5.
[196] Chow (n 14) 1300.
[197] 'Murdered Ohio Grandfather's Family Sues Facebook for Not Detecting Killer's Intent' <https://www.nbcnews.com/tech/tech-news/murdered-ohio-grandfather-s-family-sues-facebook-not-detecting-killer-n843371> accessed 27 July 2020.

day.[198] Also, content prohibition is not an easy job in every case, as can be seen in ECHR cases, where even courts fail when prohibiting content through being overreactive. Extended enforcement privatisation of hate speech risks over-affecting users' freedom of speech of users.[199]

Conclusion

> "The extent to which Facebook posts and messages have led to real-world discrimination and violence must be independently and thoroughly examined."[200]

The definition of illegal hate speech is still not formally and jointly defined. Governments, international institutes and online platforms have different definitions. This unclarity results in challenges for online platforms who have to comply with different legislations. In this paper, phrases that aim to continue or increase social disorder against minority groups such as gender, race, and religion in the society and that encourage violence are accepted as speeches of hate.

ECD offers us the fairest method to hold online platforms liable for hate speech generated by users. OPs are not liable if they are not informed or they do not actively take part in the production of the content or reaching more people. However, if they are aware of the content or if they are promoting that content, such as receiving advertisements, their liability will begin. CJEU's active/passive diligence to the issue of liability came into conflict with the recommendations of the commission and the development of the AVMSD. In other words, online platforms that have to take proactive measures may become unable to benefit from

[198] 'Underpaid and Overburdened: The Life of a Facebook Moderator | Facebook | The Guardian' <https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content> accessed 27 July 2020.
[199] Casarosa (n 10) 398.
[200] Human Rights Council (n 1).

liability immunity as they are now in an active role. OPs should be able to take responsibility by enforcing steps toward illegal activity on their platforms.[201] Maybe this uncertainty will be overcome by creating an intermediate area of liability. However, if this intermediate area is left under the responsibility of the member states, more than one law will arise, which will create challenges to comply with for online platforms. The platform, which cannot benefit from the exemption of responsibility of the ECD, may also be considered except in accordance with the legislation of the state.[202]

Users have to read Terms of Services made up of huge pages in order to see their rights and responsibilities while using online platforms. And since many of these texts are located in different links, many clicks are required.[203] Social platforms try to cope with hate speech by using automatic moderation systems. This method is achieved by implementing various algorithms that change from platform to platform. It is doubtful to what extent these algorithms can cause a sensation. If the platforms and algorithms only accept flagging materials as hate speech, a busy environment will begin on the Internet and minorities will have difficulty expressing themselves. For this reason, the right to disturb the society, which we can call a manifestation of freedom of expression, which is actually one of the most basic human rights, will be taken from people.

OPs are considered to have made commitments regarding content moderation methods. However, difficulties have been found in assessing the extent to which these commitments were fulfilled.[204] In fact according to De Streel's[205] survey, online platforms reported that their content moderation methods help to decrease digital aggressiveness and illicit content amount.

[201] Stalla-Bourdillon and Thorburn (n 99) 154.
[202] De Streel and others (n 21) 23.
[203] Tan (n 150) 100.
[204] De Streel and others (n 21) 16.
[205] De Streel and others (n 21).

At the same time, though, the innovative nature of offenders may allow them to find methods to circumvent content moderation methods. [206] The mechanisms used by online platforms in order to regulate illicit content is not sufficiently successful to safeguard universal human rights. [207] Methods for Content recognition have significant shortcoming such as lack of transparency which only exist because of the complexity of the way in which the tools work, the danger of over-control and sufficient procedural safeguards.[208] Some NGO's called for a different 'Notice-and-takedown' strategy by modifying or increasing the usage of takedown for specific types of content.[209]

Keeping in mind that completely holding the online platforms liable for the users' content will further increase the censorship on the Internet. The online platform should not be held liable because of the users' content while they are not informed about it. But platforms that monetize their users' content have a job to make the Internet a less hateful place. While performing this task, they are obliged to continually inform the states, individuals and non-governmental organizations and share their algorithms. In this way, we will be able to avoid over-control and excessive obstruction.

[206] Rozenfeldova and Sokol (n 3) 867.
[207] De Streel and others (n 21) 45.
[208] Thomas Riis and Sebastian Felix Schwemer, 'Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation' [2018] SSRN Electronic Journal 13 <https://www.ssrn.com/abstract=3300159> accessed 5 August 2020.
[209] TILT (2016), Role of online intermediaries: Summary of the public consultation, Study for the European Commission.

Bibliography

International Treaties and European Union Legislations

OHCHR | International Convention on the Elimination of All Forms of Racial Discrimination' <https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx> accessed 7 August 2020

'OHCHR | International Covenant on Civil and Political Rights' <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx> accessed 30 July 2020

The Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended).

The Additional Protocol to the Convention on Cybercrime Treaty No.189 (*CoE Treaty Office*) <https://www.coe.int/en/web/conventions/full-list> accessed 30 July 2020

Charter of Fundamental Rights of the European Union 2012

Directive 2000/31/EC of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market [2000] OJ L 178

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)

The EU Code of Conduct on Countering Illegal Hate Speech Online (*European Commission - European Commission*) <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en> accessed 28 July 2020

Case Law

Delfi As v Estonia [2015] The European Court of Human Rights Application no. 64569/09

Eva Glawischnig-Piesczek v Facebook Ireland [2018]

Eva Glawischnig-Piesczek v Facebook Ireland Limited [2019]

Gündüz v Turkey [2004] European Court of Human Rights Application no. 35071/97

Jane Doe No 1 v BackpageCom, LLC [2016] United States Court of Appeals For the First Circuit 817 F.3d 12 (1st Cir. 2016)

Laurence Godfrey v Demon Internet Limited [2001] QB [1999] EWHC QB 244

Packingham v North Carolina [2017] Suprem Court of United States 137 S. Ct. 1730 (2017) 198 L. Ed. 2d 273

Cohen v Facebook, Inc [2017] United States District Court, ED New York 252 F. Supp. 3d 140 (E.D.N.Y. 2017)

Secondary Sources: Books, Book Chapters, Academic Articles, Working Papers and Websites

'Advertiser-Friendly Content Guidelines - YouTube Help' <https://support.google.com/youtube/answer/6162278> accessed 14 August 2020

Afori OF, 'Online Rulers as Hybrid Bodies: The Case of Infringing Content Monitoring' (2020) 23 University of Pennsylvania Journal of Constitutional Law 1

Alkiviadou N, 'Hate Speech on Social Media Networks: Towards a Regulatory Framework?' (2019) 28 Information & Communications Technology Law 19

Arun C, 'Making Choices: Social Media Platforms and Freedom of Expression Norms' [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3411878> accessed 6 July 2020

Bard P and Bayer J, 'Hate Speech and Hate Crime in the EU and the Evaluation of Online Content Regulation Approaches' (2020)

Bayer J, 'Between Anarchy and Censorship' (2019) 3 CEPS Papers in Liberty and Security in Europe

Brunner L, 'The Liability of an Online Intermediary for Third Party Content The Watchdog Becomes the Monitor: Intermediary Liability After' (2016) 16 Human Rights Law Review 163

Bukovská B, ‘The European Commission’s Code of Conduct for Countering Illegal Hate Speech Online’ [2019] Transatlantic Working Group

Casarosa F, ‘The European Regulatory Approach toward Hate Speech Online: The Balance between Efficient and Effective Protection’ (2020) 22 Gonzaga Journal of International Law 391

Chow ZE, ‘Evaluating the Approaches to Social Media Liability for Prohibited Speech.’ (2019) 51 New York University Journal of International Law and Politics 1293

Cobbe J, ‘Algorithmic Censorship by Social Platforms: Power and Resistance’ <https://ssrn.com/abstract=3437304>

De Streel A and others, ‘Online Platforms’ Moderation of Illegal Content Online’ (2020)

Media Pluralism and Democracy: Report. (2016) <http://dx.publications.europa.eu/10.2838/248670> accessed 29 July 2020

European Court of Human Rights, ‘Factsheet – Hate Speech’ (2020) <https://www.echr.coe.int/Documents/FS_Hate_speech_ENG.pdf>

Facebook, ‘Community Standards’ <https://www.facebook.com/communitystandards/hate_speech> accessed 29 July 2020

‘Framework Convention for the Protection of National Minorities - European Treaty Series - No:157’ (*CoE Treaty Office*) <https://www.coe.int/en/web/conventions/full-list> accessed 30 July 2020

Frosio G, ‘The Death of “No Monitoring Obligations”: A Story of Untameable Monsters’ (2017) 8 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law 199

Frosio G and Husovec M, ‘Accountability and Responsibility of Online Intermediaries’ in Giancarlo Frosio (ed), Giancarlo Frosio and Martin Husovec, *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) <http://oxfordhandbooks.com/view/10.1093/oxfordhb/9780198837138.001.0001/oxfordhb-9780198837138-e-31> accessed 3 August 2020

Frosio GF, ‘Reforming Intermediary Liability in the Platform Economy: A European Digital Single Market Strategy’ (2017) 4 Centre for International Intellectual Property Studies Research Paper

Geiger C, Frosio G and Izyumenko E, ‘Intermediary Liability and Fundamental Rights’ (Center for International Intellectual Property Studies 2019) 6

‘Godfrey v. Demon Internet Limited’ (*Global Freedom of Expression*) <https://globalfreedomofexpression.columbia.edu/cases/godfrey-v-demon-internet-limited/> accessed 31 July 2020

‘Google to Exclude “revenge Porn” from Internet Searches | Google | The Guardian’ <https://www.theguardian.com/technology/2015/jun/20/google-excludes-revenge-porn-internet-searches> accessed 3 August 2020

Great Britain and others, *Online Harms White Paper* (2019)

‘Hate Speech Policy - YouTube Help’ <https://support.google.com/youtube/answer/2801939?hl=en> accessed 29 July 2020

‘How Content ID Works - YouTube Help’ <https://support.google.com/youtube/answer/2797370?hl=en> accessed 3 August 2020

Human Rights Council, ‘Report of The Independent International Fact-Finding Mission on Myanmar’ (2018) UN Doc A/HRC/39/64

Keller D, ‘Toward a Clearer Conversation About Platform Liability’ (Knight First Amendment Institute’s 2018)

Klonick K, ‘The New Governors: The People, Rules, and Processes Governing Online Speech’ (2017) 6 Harvard Law Review 131

Koebler J and Cox J, ‘Here’s How Facebook Is Trying to Moderate Its Two Billion Users’ <https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works> accessed 12 August 2020

Kuczerawy A, ‘General Monitoring Obligations: A New Cornerstone of Internet Regulation in the EU?’, *Rethinking IT and IP law: celebrating 30 years CiTiP* (2019)

Kuklis L, ‘Video-Sharing Platforms In AVMSD – A New Kind Of Content Regulation’, *Research Handbook on EU Media Law and Policy* (Elgar Publishing 2020)

Li T, ‘Beyond Intermediary Liability: The Future of Information Platforms - Workshop Report’ [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3392438> accessed 7 July 2020

Lynskey O, ‘Regulating “Platform Power”’ [2017] SSRN Electronic Journal <http://www.ssrn.com/abstract=2921021> accessed 31 July 2020

Mac Síthigh D, ‘The Road to Responsibilities: New Attitudes towards Internet Intermediaries’ (2020) 29 Information & Communications Technology Law 1

Madiega T and others, *Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In-Depth Analysis.* (2020) <https://op.europa.eu/publication/manifestation_identifier/PUB_QA0420239ENN> accessed 31 July 2020

Montagnani ML, ‘A New Liability Regime for Illegal Content in the Digital Single Market Strategy’ (Bocconi University 2019) 477007

Mostert F, ‘Free Speech and Internet Regulation’ (2019) 14 Journal of Intellectual Property Law & Practice 607

‘Murdered Ohio Grandfather’s Family Sues Facebook for Not Detecting Killer’s Intent’ <https://www.nbcnews.com/tech/tech-news/murdered-ohio-grandfather-s-family-sues-facebook-not-detecting-killer-n843371> accessed 27 July 2020

‘New Education Programme Launched to Combat Online Piracy’ (*GOV.UK*) <https://www.gov.uk/government/news/new-education-programme-launched-to-combat-online-piracy> accessed 3 August 2020

‘OHCHR | Governments and Internet Companies Fail to Meet Challenges of Online Hate – UN Expert’ <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25174&LangID=E> accessed 9 August 2020

Patty M, ‘Social Media and Censorship: Rethinking State Action Once Again’ (2019) 40 Mitchell Hamline Law Journal of Public Policy and Practice 99

Quintel T and Ullrich C, ‘Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond’ in BILYANA PETKOVA and TUOMAS OJANEN (eds), *Fundamental Rights Protection Online: the Future Regulation of Intermediaries* (Edward Elgar Publishing 2019)

‘Regulating Content Moderation in Europe beyond the AVMSD’ (*Media@LSE,* 25 February 2020) <https://blogs.lse.ac.uk/medialse/2020/02/25/regulating-content-moderation-in-europe-beyond-the-avmsd/> accessed 9 August 2020

Reynders D, ‘5th Evaluation of the Code of Conduct’ 5

Riis T and Schwemer SF, ‘Leaving the European Safe Harbor, Sailing Towards Algorithmic Content Regulation’ [2018] SSRN Electronic Journal <https://www.ssrn.com/abstract=3300159> accessed 5 August 2020

Rozenfeldova L and Sokol P, ‘Liability Regime of Online Platforms New Approaches and Perspectives’ (2019) 3 EU and Comparative Law Issues and Challenges Series 866

Sally Broughton Micova, ‘The Audiovisual Media Services Directive: Balancing Liberalisation and Protection’ in Elda Brogi and Pier Luigi Parcu (ed), *Forthcoming Handbook on EU Media Law and Policy* (Edward Elgar Publishing 2020) <https://ssrn.com/abstract=3586149>

Savin A, ‘New Directions in EU Policymaking on the Content Layer: Disruption and Law’ (Copenhagen Business School Law 2020) 20–05

Stalla-Bourdillon S, ‘Internet Intermediaries as Responsible Actors? Why It Is Time to Rethink the E-Commerce Directive as Well’ in Mariarosaria Taddeo and Luciano Floridi (eds), *The Responsibilities of Online Service Providers*, vol 31 (Springer International Publishing 2017) <http://link.springer.com/10.1007/978-3-319-47852-4_15> accessed 7 July 2020

Stalla-Bourdillon S and Thorburn R, ‘The Scandal of Intermediary: Acknowledging the Both/and Dispensation for Regulating Hybrid Actors’ in B.Petkova and T.Ojanen (ed),

Fundamental Rights Protection Online: the Future Regulation of Intermediaries (Edward Elgar Publishing 2019)

Stuart AH, 'Social Media, Manipulation, and Violence' (2019) 15 South Carolina Journal of International Law and Business 100

Tan C, *Regulating Content on Social Media: Copyright, Terms of Service and Technological Features* (UCL Press 2018) <http://www.jstor.org/stable/10.2307/j.ctt2250v4k> accessed 14 August 2020

'The Glawischnig-Piesczek v Facebook Case: Knock, Knock. Who's There? Automated Filters Online' (*CITIP blog*, 12 November 2019) <https://www.law.kuleuven.be/citip/blog/the-glawischnig-piesczek-v-facebook-case-knock-knock-whos-there-automated-filters-online/> accessed 5 August 2020

'Twitter Terms of Service' <https://twitter.com/tos?lang=en#usContent> accessed 7 August 2020

Tworek H and Leerssen P, 'An Analysis of Germany's NetzDG Law' (Transatlantic High Level Working Group 2019)

'Underpaid and Overburdened: The Life of a Facebook Moderator | Facebook | The Guardian' <https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content> accessed 27 July 2020

UNESCO, 'Freedom of Expression: A Fundamental Human Right Underpinning All Civil Liberties' (*UNESCO*, 14 April 2015) <https://en.unesco.org/70years/freedom_of_expression> accessed 29 July 2020

'United Nations Office on Genocide Prevention and the Responsibility to Protect' <https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml> accessed 30 July 2020

'Using Technology to Protect Intimate Images and Help Build a Safe Community' (*About Facebook*, 5 April 2017) <https://about.fb.com/news/2017/04/using-technology-to-protect-intimate-images-and-help-build-a-safe-community/> accessed 3 August 2020

‘Visa, MasterCard Move To Choke WikiLeaks’ <https://www.forbes.com/sites/andygreenberg/2010/12/07/visa-mastercard-move-to-choke-wikileaks/> accessed 3 August 2020

Wenguang Y, ‘Internet Intermediaries’ Liability for Online Illegal Hate Speech’ (2018) 13 Frontiers of Law in China 342

Legislative Reports and Declarations

Committee of Ministers, ‘Recommendation No. R (97) 20 of The Committee Of Ministers To Member States on "Hate Speech’ (1997)

‘Declaration on Freedom of Communication on the Internet’ (2003) 840th meeting of the Ministers' Deputies

European Commission, ‘Online Platforms and the Digital Single Market. Opportunities and Challenges for Europe’ (2016) Communication COM(2016) 0288 final

European Commission, ‘Tackling Illegal Content Online. Towards an Enhanced Responsibility of Online Platforms.’ (2017) Communication COM(2017) 555

European Commission, ‘Recommendation 2018/334 on Measures to Effectively Tackle Illegal Content Online’ (2018) OJ [2018] L 63/50