Balancing Content Moderation and Human Rights in the Digital Age: Analyzing the Trusted Flaggers Mechanism and the Responsibilities of the Online Platforms under the Digital Services Act

Bengü Özer

(Student number: 2101363)

Thesis supervisor: Mr. Dr. Colette Cuijpers

Second reader: Dr. Jasper van den Boom

Academic year: 2022-2023

Tilburg Institute for Law, Technology, and Society (TILT)

Master Law & Technology

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my supervisors, Mr. Dr. Colette Cuijpers and Dr. Jasper van den Boom. Their guidance, expert insights, support, and encouragement have been invaluable in this academic journey.

I extend my heartfelt thanks to my family members, Hidayet Özer, Necla Özer, and Taylan Can Özer. Their boundless support, both emotionally and morally, has been the cornerstone of my perseverance.

Lastly, I wish to offer a special acknowledgment to my dear friends and colleagues, Sefa Geçikli and Cem Erbay. Their meticulous proofreading of my thesis and unwavering encouragement throughout this journey have not only made it more manageable but also enjoyable.

Tilburg,

21st August 2023

CHAPTER I - INTRODUCTION

1.1. Background and Problem Statement

In 2022, a new set of rules in EU came into force known as the Digital Service Act (DSA)[1]. The scope and ambit of the DSA is to ensure a safe and transparent online environment. The DSA is primarily focused upon the online marketplace, regulating the responsibilities of digital services that act as intermediaries in their position of connecting customers with goods, services, and content.[2] According to the European Commission (EU) the DSA will provide better protection to consumers and their fundamental rights online and establish a single framework across the EU while creating a transparent and accountable framework for online platforms.[3]

Since one of the major goals of the DSA is tackling the spread of illegal content online, it defines clear responsibilities for "providers of intermediary services, and in particular online platforms"[4]. Chapter III of the act contains due diligence obligations for a transparent and safe online environment in five sub-sections.[5] The third one under Article 22 DSA introduces a new instrument called "trusted flaggers". The definition of this instrument is within the context of "notice-and-action"[6] procedures established by online platforms for tackling illegal content. Pursuant to Recital 46 of the DSA, "*Action against illegal content can be taken more quickly and reliably where online platforms take the necessary measures to ensure that notices submitted by trusted flaggers through the notice and action mechanisms required by this Regulation are treated with priority"* and *"Such trusted flagger status should only be awarded to entities, and not individuals."*[7] It can be deduced from the recital that these trusted flaggers cannot be individuals but entities that have demonstrated particular expertise in identifying illegal content and must be independent of any provider of online platforms[8] such as NGOs[9] and they shall be given priority in terms of the notices of detecting illegal content.

---

[1]Regulation on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [DSA] [2022] OJ L 227/1

[2] European Commission, 'Digital Services Act: EU's landmark rules for online platforms enter into force' (2022) < https://ec.europa.eu/commission/presscorner/detail/en/IP_22_6906> accessed on 21 November 2022

[3]European Commission, 'Questions and Answers: Digital Services Act' (2022) <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348> accessed on 20 November 2022

[4] *DSA*

[5] *Ibid*

[6] Digital Services Act, Article 14

[7] Digital Services Act, Recital 46

[8]Claude-Étienne Armingaud, Dr. Ulrike Elteste, Camille J. Scarparo, Dr. Thomas Nietsch, Andreas Müller 'EU Digital Services Act: Fundamental Changes for Online Intermediaries' (11 April 2022)

The DSA states that the status of trusted flaggers shall be awarded to entities upon application by the Digital Services Coordinator ('DSC')[10] of the member state, in which the applicant is established (art. 22 and 3(n) DSA)[11]. The question which arises before us is what are the criteria to qualify as a "trusted flagger"? What is the assessment model that aids the DSC to decide who is a "trusted flagger"? Certainly, the letter of the law provides us with an answer under Article 22(2) that the entity has to demonstrate that it is independent from any provider of online platforms[12] and has expertise for detecting, identifying and notifying illegal content[13] and while doing so the entity has to carry out these activities for the purposes of submitting notices diligently, accurately and objectively[14]. However, at this juncture the ambiguity arises, how does DSC decide the expertise and competence of the entity? How does the entity in its application demonstrate independence? And lastly, how does one describe the notion of diligence, accuracy of submitting notices? These questions are important to delve into as the very existence and essence of the DSA relies upon the cooperation of trusted flaggers and therefore the DSA falls apart if the foundation of the "trusted flaggers" is tremulous.

In this digital age, content moderation may be the need of the hour but how can content be moderated in the light of the DSA, when there is ambiguity not cleared by authorities that how is a "flagger" considered to be "trusted" after all? It is pertinent to note that the vagueness on how to obtain and regulate this new concept of "trusted flaggers" brings other problems like transparency and accountability[15]. In order to create such flagging mechanism as stated under Article 16, DSA are required to be adapted to identify and eliminate illegal content as fast[16] as possible[17], using algorithmic tools will be required.[18]

---

https://www.klgates.com/eu-digital-services-act-fundamental-changes-for-online-intermediaries-11-4-2022 accessed 13 March 2023

[9]Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021 http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022

[10] Digital Services Act, Article 22

[11] Digital Services Act, Article 3(n)

[12] Digital Services Act, Article 22(b)

[13] Digital Services Act, Article 22(a)

[14] Digital Services Act, Article 22(c)

[15]Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021 http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022

[16] Digital Services Act, Article 16

[17] Communication from The Commission to The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions Tackling Illegal Content Online towards an enhanced responsibility of online platforms (2017) COM/2017/0555 final

[18] Joseph Downing 'The EU's Digital Services Act: Europeanising social media regulation?' (2022) LSE Comment https://blogs.lse.ac.uk/europpblog/2022/08/08/the-eus-digital-services-act-europeanising-social-media-regulation/ accessed 23 November 2022

Using such tools is criticized for not being transparent enough[19] on how to decide what is illegal as well as missing the actual harmful content because of the complex nature of the creations online[20], as being the "unaccountable" part of the problem. In addition to that, this automated content moderation might also "damage user experience by over-detection and the generation of false-positives" and "false negatives".[21] Therefore it cannot be deemed as a full-functioning solution for flagging alone and will require human involvement,[22] which will lead to huge economic investments by the platforms.[23] This thesis will focus not on the technical aspects of algorithm usage to detect and remove illegal content, but the analysis of building an effective system for tackling illegal content and possible outcomes of content moderation in the context of freedom of expression.

The Charter of Fundamental Rights of the European Union ('CFR') states that regardless of the medium, everyone has a right to have a freedom to hold opinions, receive and impart information and ideas without unwanted intervention of the public authority[24]. Another key issue therefore arises in the context of online content moderation which comes along as a result of the aforementioned over-detection or over-removal of user generated creations, which may result in violation of freedom of speech and expression.[25] The DSA sets rules governing moderation and removal of illegal content from online platforms by the hands of trusted flaggers as a part of the notice-and-action mechanism, however, without sufficiently explaining the safeguards assuring the right to free speech and expression.[26] This way, it also raises concerns regarding the DSA causing digital censorship[27] on the online

[19] Joan Donovan 'Why Social Media Can't Keep Moderating Content in the Shadows' (2020) MIT Technology Review https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/ accessed 24 November 2022

[20] Joseph Downing 'The EU's Digital Services Act: Europeanising social media regulation?' (2022) LSE Comment https://blogs.lse.ac.uk/europpblog/2022/08/08/the-eus-digital-services-act-europeanising-social-media-regulation/ Accessed 23 November 2022

[21] IMCO Committee, Online Platforms' Moderation of Illegal Content (2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf> accessed 21 November 2022

[22] *Ibid*

[23] *Ibid*

[24] Charter of Fundamental Rights of the European Union [2012] OJ C 326, Article 11

[25] Valentina Golunova, The Digital Services Act and freedom of expression: triumph or failure? (2021) Maastricht University Blog <https://www.maastrichtuniversity.nl/blog/2021/03/digital-services-act-and-freedom-expression-triumph-or-failure> accessed 22 November 2022

[26] Brussels Report, 'The DSA is nothing more than digital censorship' (2022) https://www.brusselsreport.eu/2022/01/20/the-dsa-is-nothing-more-than-digital-censorship/ accessed 23 November 2022

[27] *Ibid;* Decision no. 2020-801 DC 'Law to combat hateful content on the internet' https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm accessed 23 November 2022

platforms and posing a threat to fundamental rights while trying to provide a transparent and accountable online environment.

With the increasing importance of online platforms over the years, market players have become an important part of content moderation and enforcement. Especially social media companies are often criticized for failing to take down harmful content, and for when they actually do so.[28] While some argue that social media censorship infringes on free speech rights[29], others believe that it is necessary to prevent harm. The shift from state to market players in content moderation over time has complicated the tension between freedom of expression and content moderation on digital platforms. How these platforms will cooperate with trusted flaggers is also an integral part of this whole issue of shift.

1.2. Literature Review

The literature on what is the purpose of trusted flaggers and how they work, contains conflicted views. There are, to begin with, debates on whether they are necessary for protection against illegal and harmful content online[30] or leave power in the hands of tech giants arbitrarily.[31] Some argue that a balanced removal policy can be implemented with trusted flaggers[32] for a safer digital space, while others strongly point out that it will only lead to digital censorship and surveillance[33] culture and harm the essence of right to free speech.[34]

---

[28] United Nations Human Rights Office of the High Commisioner, 'Moderating online content: fighting harm or silencing dissent?', 23 July 2021, https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent> accessed 12 March 2023

[29] Will Oremus, 'How social media 'censorship' became a front line in the culture war',9 October 2022, <https://www.washingtonpost.com/technology/2022/10/09/social-media-content-moderation/> accessed 12 March 2023

[30] Christopher Herwartz 'Now there's finally an answer to the destructive power of social media' https://www.handelsblatt.com/meinung/kommentare/kommentar-jetzt-gibt-es-endlich-eine-antwort-auf-die-zerstoererische-kraft-der-sozialen-medien/27981470.html?ticket=ST-1888944-JCAOkcZgsvjjp1MOpff4-ap1 accessed 24 November 2022

[31] Brussels Report, 'The DSA is nothing more than digital censorship' (2022) https://www.brusselsreport.eu/2022/01/20/the-dsa-is-nothing-more-than-digital-censorship/ accessed 23 November 2022

[32] Teresa Rodríguez de las Heras Ballell, 'The background of the Digital Services Act: looking towards a platform economy Teresa Rodríguez de las Heras Ballell' ERA Forum (2021) https://link.springer.com/article/10.1007/s12027-021-00654-w accessed 22 November 2022

[33] Sebastian Becker Castellaro & Jan Penfrat, 'The DSA fails to reign in the most harmful digital platform businesses – but it is still useful' , Verfassungsblog on Matters Constitutional, (2022) https://verfassungsblog.de/dsa-fails/ accessed 24 November 2022

[34] Brussels Report, 'The DSA is nothing more than digital censorship' (2022) https://www.brusselsreport.eu/2022/01/20/the-dsa-is-nothing-more-than-digital-censorship/ accessed on 23 November 2022

In the European Union, liability of the providers of intermediary services for the online content created by its users is regulated by the e-Commerce Directive ('ECD')[35] until the DSA entered into force. The principle in the ECD was that providers of intermediary services are not liable for the content propagated through their platforms or created by their users, given that they did not actively contribute to the propagating, or eliminated the illegal content after they received information.[36] The DSA also lets the providers of intermediary services enjoy the same exemption of liability, however, different than the ECD, it brings further due diligence obligations to these providers[37].

Hosting services are "where an information society service is provided that consists of the storage of information by the end user" according to Article 6 of the DSA.[38] Online platforms are the hosting services that not only store but also propagate information upon the request of the user, which makes the online content available to infinite number of third parties.[39] Providers of the 'hosting services' shall benefit some liability exemptions if they act diligently for detecting and removing illegal online content under the DSA just as in the e-Commerce Directive.[40] Article 7 of the DSA, the so-called Good Samaritan clause[41], lets providers of intermediary services enjoy exemption from liabilities "solely because they carry out voluntary own-initiative investigations or other activities aimed at detecting, identifying and removing, or disabling of access to, illegal content, or take the necessary measures to comply with the requirements of Union law, including those set out in this Regulation".[42] This provision requires increasing involvement of these providers of intermediary services, and specifically online platforms, to be more active at tackling illegal content, but it also creates the risk of removing more than the necessary amount of content to avoid liability and

[35] Aleksandra Kuczerawy, 'Intermediary liability & freedom of expression: Recent developments in the EU notice & action initiative' (2015) https://www-sciencedirect-com.tilburguniversity.idm.oclc.org/science/article/pii/S0267364914001836?fr=RR-2&ref=pdf_download&rr=76f52ff358671c8a accessed 21 November 2022

[36] Sally O'Brien, 'E-Commerce Directive versus the new Digital Services Act: is there a new liability regime for online service providers?' (2022) https://www.loganpartners.com/e-commerce-directive-versus-the-new-digital-services-act-is-there-a-new-liability-regime-for-online-service-providers/ accessed on 25 November 2022

[37] Chapter III of Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277

[38] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277

[39] Christian Alberdingk Thijm, '10 Questions about the Digital Services Act, 2022, https://www.bureaubrandeis.com/10-questions-about-the-digital-services-act/ accessed 21 November 2022

[40] Aleksandra Kuczerawy, 'Intermediary liability & freedom of expression: Recent developments in the EU notice & action initiative' (2015) https://doi.org/10.1016/j.clsr.2014.11.004 accessed 21 November 2022

[41] *Ibid*

[42] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277

harm users' right to freedom of expression at the same time.[43] How exactly the cooperation between these platforms and trusted flaggers[44] to counter illegal content online takes place while mitigating abovementioned risks is not quite clear and this is the gap in existing research this thesis aims to fill.

1.3. Research Purpose and Questions

In my thesis, I will focus how to strike a balance between implementing trusted flagging mechanism system in content moderation and protecting the freedom of expression by analyzing the relevant provisions of the DSA. I aim to provide meaningful contribution to the current heated debates on the possible drawbacks of this mechanism as well as proposing safeguards to protect the essence of the fundamental human rights while tackling illegal content online.

The existing literature on content moderation and freedom of expression online has focused on the challenges and criticisms of the current mechanisms for tackling illegal content, and the privatization of enforcement by online platforms. However, there is a lack of empirical and theoretical studies on the impact and effectiveness of the trusted flaggers mechanism introduced by the DSA, which aims to provide a more transparent and accountable framework for online platforms. My thesis will contribute to this gap by analyzing how the trusted flaggers mechanism works, what are the advantages and disadvantages of this system for content moderation and fundamental human rights, and what are the possible safeguards to protect competing interests while creating a safer digital environment.

My thesis will answer the following question "How does the implementation of trusted flaggers system under the DSA affect the freedom of expression within the context of online content moderation?"

The sub-questions which are going to be covered in this thesis are as follows:

1- What obligations are laid down in the DSA for content moderation and what does the introduction of the DSA change for platform operators?

---

[43] Aleksandra Kuczerawy, 'The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act' Verfassungsblog on Matters Constitutional (2021) https://verfassungsblog.de/good-samaritan-dsa/ accessed 23 November 2022

[44] European Commission, 'Questions and Answers: Digital Services Act' (2022) https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 accessed 20 November 2022

2- What are the concerns surrounding the use of trusted flagger systems for content moderation and what are its advantages and disadvantages?

3- How can the trusted flagger system be implemented to promote content moderation that is effective in removing illegal content, yet proportional with respect to the freedom of expression?

1.4. Methodology & Methods

To answer to the main research question and the sub-questions, doctrinal legal research will be conducted. Because the objective of this thesis is to address the question of how to strike a balance between content moderation and fundamental human rights while analyzing the trusted flagging mechanism; the role of online platforms regarding moderating online content, and challenges as well as how these concerns can possibly be addressed to ensure transparency and accountability as set in the DSA will be analyzed. The provisions from other legislations regarding content moderation and tackling illegal content will also be included and compared to the ones in the DSA in order to see what is novel.

Relevant case law regarding the interplay between online content moderation and freedom of expression[45] along with the other fundamental rights will also be used in this thesis to understand the scope of the issue and the courts' views while making decisions. Relevant provisions in the DSA, the ECD and EU case law will be the primary sources for this research while the works of legal scholars will be used as the secondary sources. Moreover, legal literature focusing on the ratio behind introducing trusted flaggers in relation with the content moderation will be reviewed in detail by comparing the different methods of content moderation with the involvement of trusted flaggers[46] to substantiate the arguments made by legal scholars.

1.5. Chapter Structure

The first chapter provides background information on the major issues regarding content moderation and how trusted flaggers work in the notice-and-action system. Following this chapter, the second chapter of this study will provide an analysis regarding changes in content moderation by comparing the DSA with the ECD. The primary challenges and the liability regime for online platforms will also be discussed in this chapter.

---

[45]

[46] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

After comparing the DSA and the ECD in terms of the liabilities of the online platforms in content moderation in the second chapter, what exactly trusted flaggers are according to the DSA and how they will work in content moderation, its advantages and disadvantages along with the possible legal challenges they pose will be discussed in the third chapter.

The fourth and the last chapter will delve into the implementation of trusted flagger systems for content moderation, focusing on their effectiveness, proportionality. Concurrently, recommendations will be proposed for efficient use of trusted flagging mechanism for content moderation while protecting right to freedom of expression. This will encompass exploring different approaches to implementing trusted flagger systems, evaluating their efficacy and proportionality in removing illegal content while safeguarding freedom of expression, and discussing recommendations and suggestions for enhancing these systems and striking a balance between competing interests.

Finally, in the conclusion, I will provide a comprehensive analysis of how trusted flaggers system can effectively contribute to fostering a safer and more accountable online environment in the light of the DSA. This analysis will also include the influence of trusted flaggers on content moderation practices of platform operators while upholding the fundamental principles of freedom of expression.

CHAPTER II - CONTENT MODERATION IN THE DIGITAL AGE: HOW THE DSA RESHAPES THE ROLE OF PLATFORM OPERATORS

2.1. Chapter Introduction

The DSA is a comprehensive regulatory framework that stipulates the responsibilities of digital services that function as intermediaries in the European Union (EU) to facilitate access to goods, services, and content. "Digital services" in this context refers to online platforms that provide services such as social media networks and marketplaces.[47] It establishes due diligence obligations for online platforms and other online intermediaries to comply with. For instance, under the new rules the users can flag illegal content, and also have a mechanism of contesting content moderation practices of platforms.[48] It aims to create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses[49] by imposing more transparency requirements for online platforms regarding their decisions on content removal and moderation.[50]

The DSA updates and complements the current EU framework for digital services, the ECD[51], which establishes harmonized rules on issues like transparency and liability of online services since its enactment in 2000.[52] The ECD is still the foundation of digital regulation, but the online environment has transformed since it was enacted, and the DSA aims to tackle these changes and challenges especially in relation to online intermediaries.[53] The rules applied under the ECD are being changed in order to regulate digital services while utilizing *their operational and technical capacity to tackle illegal content*[54] and protecting fundamental rights.[55] Therefore, it is important to understand how this new set of rules addresses these challenges and what implications this has for digital services and users by analyzing both the DSA and the ECD in the context of content moderation.

---

[47] European Commission, 'Digital Services Act: Questions and Answers' (24 April 2023) https://digital-strategy.ec.europa.eu/en/faqs/digital-services-act-questions-and-answers accessed 10 May 2023
[48] *Ibid*
[49] *Ibid*
[50] *Ibid*
[51] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') OJ L 178, 17.7.2000
[52] European Commission, 'e-Commerce Directive' https://digital-strategy.ec.europa.eu/en/policies/e-commerce-directive accessed 10 May 2023
[53] *Ibid*
[54] Digital Services Act, Recital 27 "their technical and operational ability to act against specific items of illegal content"
[55] Berrak Genç-Gelgeç, 'Regulating Digital Platforms: Will the DSA Correct its Predecessor' s Deficiencies?' (2022) 18 CYELP 25 https://www.cyelp.com/index.php/cyelp/article/view/485 accessed 9 August 2023

In section 2.2, the liability regime for the online platforms under the ECD will be discussed to demonstrate the changes under the DSA in section 2.3. After that, section 2.4 will conclude the legal shift in content moderation from the ECD to the DSA by outlining the previous sections.

2.2. How was it under the E-Commerce Directive?

The ECD is the main legal framework for the provision of digital services in the EU. It regulates content moderation by online platforms in several ways. The introduction of the ECD was accompanied by a debate that took into account various factors such as the emerging nature of the digital sector, technical capabilities of online intermediaries, and changes in the online environment.

The directive aimed to remove obstacles to cross- border online services and establish harmonized rules on transparency, information requirements, commercial communications, electronic contracts, and limitations of liability for intermediary service providers.[56] The digital sector was rapidly evolving at the time of introducing the ECD, and there was a recognized need for a legal framework to address and accommodate this changing landscape of online services. It aimed to strike a balance between the responsibilities of online intermediaries and the protection of fundamental human rights, such as freedom of expression.[57]

First, it establishes a liability regime that exempts online platforms from liability for illegal content hosted by them, as long as they do not have actual knowledge[58] of it and act expeditiously to remove[59] or disable access to it upon obtaining such knowledge. Liability of Intermediary Service Providers is regulated under Section 4 of the ECD in the articles 12, 13 and 14.[60] The limitations on liability provided for by the ECD are established horizontally,

[56] European Commission, 'e-Commerce Directive' Shaping Europe's Digital Future, June 2022, https://digital-strategy.ec.europa.eu/en/policies/e-commerce-directive accessed 11 August 2023

[57] Alexandre De Streel & Martin Husovec, 'The e-commerce Directive as the cornerstone of Internal Market: Assessment and Options for Reform' European Parliament Policy Department for Economic, Scientific and Quality of Life Policies May 2020, https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)648797 accessed at 11 August 2021

[58] CMS Law Now, 'The E-commerce Directive' (12 April 2002) https://cms-lawnow.com/en/ealerts/2002/04/the-e-commerce-directive?format=pdf&v=4 accessed 10 May 2023

[59] Alexandre De Streel et al., Center on Regulation in Europe, 'Online Platforms' Moderation of Illegal Content Online' (June 2020) https://cerre.eu/news/study-online-platforms-moderation-of-illegal-content-online/ (accessed 10 May 2023)

[60] Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') OJ L 178, 17.7.2000

which means that they cover liability, for any kind of illegal activities initiated by third parties.[61]

Under article 12 of the ECD regarding "Mere Conduit", if a service provider only acts as a passive channel for information and does not interfere with it, they will not be held responsible for it.[62] This applies when the provider either sends information through communication networks or gives access to a communication network.[63] Under article 13 of the ECD regarding "Caching", a service provider who only sends information from a user through a communication network is not liable for damages caused by that information if the information is stored temporarily and automatically to make the transmission faster and more efficient for other users who request it.[64] The service provider must follow certain conditions such as not changing the information and removing access to it if the information has been deleted or blocked at its original source.[65] Last but not least, under article 14 of the ECD regarding "Hosting", a service provider who offers hosting services that store information from a user will not be liable for damages either, if the service provider does not know or have reason to know that the information was breaking any law; and quickly removes or blocks access to the information when they find out or suspect that it was unlawful.[66] Due to its inherent characteristics, hosting services are susceptible to the infiltration of illegal content uploaded by their users, and are bound by more rigorous exemption regulations in comparison to other categories of online intermediaries, such as conduit and caching service providers.[67] Henceforth, the aforementioned providers are exempted from any liabilities arising from illegal content uploaded by their users, provided that (i) they lack actual knowledge of the user's illegal activities and are not aware of any facts or circumstances that would indicate such activities or information, and (ii) upon acquiring such knowledge, they promptly ("expeditiously" in the ECD) take measures to remove or restrict access to the information.[68] The exemption in question is exclusively applicable to cases where the

[61] First Report on the application of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, COM(2003) 702, 21.11.2003
[62] Directive on Electronic Commerce, Article 12
[63] CMS Law Now, 'The E-commerce Directive' (12 April 2002) https://cms-lawnow.com/en/ealerts/2002/04/the-e-commerce-directive?format=pdf&v=4 accessed 10 May 2023
[64] *Ibid*
[65] *Ibid*
[66] *Ibid*
[67] Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in light of Recent Developments: Is it still necessary to maintain it?' JIPITEC 13 (3) 2022, https://www.jipitec.eu/issues/jipitec-13-3-2022/5555 accessed 11 May 2023
[68] *Ibid*

conduct of the hosting service providers is considered purely technical, automatic, and passive.[69] This denotes that the online intermediary possesses neither awareness nor authority over the information that is conveyed or retained.[70] A case demonstrating whether the exemption is applicable to the online intermediary in question, is *Google France, Google Inc v Louis Vuitton Malletier SA and Others*. In this case, the CJEU ruled that "Article 14 of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce… must be interpreted as meaning that the rule laid down therein applies to an internet referencing service provider in the case where that service provider has not played an active role of such a kind as to give it knowledge of, or control over, the data stored. If it has not played such a role, that service provider cannot be held liable for the data which it has stored at the request of an advertiser, unless, having obtained knowledge of the unlawful nature of those data or of that advertiser's activities, it failed to act expeditiously to remove or to disable access to the data concerned" [71] by following Recital 42 of the ECD[72], which also shows how this liability exemption is interpreted by the court.

In addition to the liability exemption, the ECD prohibits Member States from imposing a general monitoring obligation on online platforms to actively seek facts or circumstances indicating illegal activity on their services.[73] This safe harbor regime has its source in article 15(1) of the ECD which prohibits member states from requiring online intermediaries to monitor the information transmitted or stored on their services, or to actively check for facts or situations indicating illegal activity as a general obligation.[74] The enactment of the prohibition on general monitoring obligation was set by five primary rationales[75]: First, in their nascent stage, online intermediaries were initially deficient in technical capabilities to actively and accurately monitor the vast amount of information

[69] Directive on Electronic Commerce, Recital 42
[70] Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in light of Recent Developments: Is it still necessary to maintain it?' JIPITEC 13 (3) 2022, https://www.jipitec.eu/issues/jipitec-13-3-2022/5555 (accessed 11 May 2023)
[71] Case C-236/08, *Google France, Google Inc v Louis Vuitton Malletier SA and Others* [2010], ECLI:EU:C:2010:159, Ruling Paragraph
[72] *Ibid*, paras 113-116.
[73] Tambiama Madiega, European Parliamentary Research Service, 'Digital Services Act' (17 November 2022) https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689357 accessed 11 May 2023
[74] Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in light of Recent Developments: Is it still necessary to maintain it?' JIPITEC 13 (3) 2022, https://www.jipitec.eu/issues/jipitec-13-3-2022/5555 accessed 11 May 2023
[75] *Ibid*

transmitted through their platforms.[76] Second, the imposition of the monitoring requirement was deemed unjust due to its imposition of an excessive burden on passive intermediaries.[77] Third, it was deemed imperative to avoid excessive regulation that could hinder the growth of the electronic commerce industry in the EU.[78] Fourth, there was a concern about the possibility of blocking legitimate content, which could impede the free flow of information within the single market, which could occur due to false positives generated by automated systems or the tendency to avoid liability.[79] Finally, the risk of unintentionally creating knowledge and awareness of illegal content that might go undetected through general proactive monitoring was deemed to be avoided.[80] However, Recital 47 of the ECD states that this prohibition regulates the monitoring obligation in general only and does not include monitoring obligations in a specific case.[81] Moreover, it does not prevent national courts from ordering the online intermediary to take action to prevent an infringement, nor does it prevent member states from imposing a responsibility on hosting service providers to identify and prevent specific illegal activities[82] as well as aiming to undue interference with privacy and communications of users in order to protect their fundamental rights.[83]

According to the European Commission[84], the ECD's liability regime finds a middle ground among the different interests involved, especially between the interests of intermediary services, the public interest that illegal information is removed fast, and the respect of conflicting fundamental rights.[85] However, the ECD also has regulatory gaps that cause legal ambiguity in using its special liability rules.[86] It does not provide a clear definition of illegal content[87], nor does it harmonize the substantive rules on what constitutes

---

[76] *Ibid*
[77] *Ibid*
[78] *Ibid*
[79] *Ibid*
[80] *Ibid*
[81] *Ibid*
[82] *Ibid*
[83] Tambiama Madiega, European Parliamentary Research Service, 'Digital Services Act' (17 November 2022) https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689357 accessed 11 May 2023
[84] Commission Staff Working Document 'Online services, including e-commerce, in the Single Market', SEC (2011) 1641 (final) accompanying COM (2011) 942, p. 24
[85] Anja Hoffmann & Alessandro Gasparotti, 'Liability for illegal content online: Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a "Digital Services Act"' (March 2020) https://www.cep.eu/fileadmin/user_upload/hayek-stiftung.de/cepStudy_Liability_for_illegal_content_online.pdf (accessed 15 May 2023)
[86] *Ibid*
[87] European Telecommunications Network Operators' Association, 'A clean and open Internet: Public consultation on procedures for notifying and acting on illegal content hosted by online intermediaries' https://www.etno.eu/datas/positions-papers/2012/etnoc01-dsm-notice-and-action-consultation-sep-2012.pdf accessed 15 May 2023

illegality[88]. Therefore, the application and interpretation of the ECD rules on content moderation depend on the national laws[89] and courts of each Member State, as well as on the terms of service and community standards of each online platform. Such absence of harmonized rules prevents adequate protection for fundamental rights, creating legal ambiguity and disparity in an already complicated regulatory environment.[90] As a result of these challenges[91], the EU has proposed a new legislation, the Digital Services Act (DSA), which aims to address the issues of illegal content online in a more harmonized way.

2.3. What has changed under the DSA?

The DSA is a regulation that aims to create a safer and more accountable online environment for consumers and businesses in the EU.[92] The DSA is part of the Digital Services Act package, which also includes the Digital Markets Act[93] ('DMA'), a proposal to regulate the large online platforms acting as gatekeepers in the digital economy.[94]

The ECD was adopted in 2000 and established a legal framework for online services in the EU. Since then, technology has evolved and new challenges have emerged, such as the increasing dependency on online platforms to access and distribute products, services and information, the spread of illegal content online, and the impact of online platforms on fundamental rights, democracy, and public discourse.[95] Consequently, over the time, the

---

[88] Raphaël Gellert and Pieter Wolters, 'The revision of the European framework for the liability and responsibilities of hosting service providers: Towards a better limitation of the dissemination of illegal content' (7 April 2021), page 28, https://www.eerstekamer.nl/eu/documenteu/_the_revision_of_the_european/f=/vlmulhixjsbs.pdf accessed 15 May 2023

[89] *Ibid*

[90] Ilaria Buri and Joris van Hoboken, 'The Digital Services Act (DSA) proposal: a critical overview', Digital Services Act (DSA) Observatory, Institute for Information Law (IViR), University of Amsterdam, Discussion Paper (28 October 2021) https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf accessed 15 May 2023

[91] Commission Staff Working Document, 'Executive Summary of the Impact Assessment Report Accompanying the document Proposal For A 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 {COM(2020) 825 final} - {SEC(2020) 432 final} - {SWD(2020) 348 final}' https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act accessed 15 May 2023

[92] European Commission, 'Questions and Answers: Digital Services Act' (2022) https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 accessed 20 November 2022

[93] The Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC

[94] European Commission, 'The Digital Services Act Package' (2023) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed 12 August 2023

[95] European Commission, 'The Digital Services Act: ensuring a safe and accountable online environment: What are the key goals of the Digital Services Act?' https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#what-are-the-key-goals-of-the-digital-services-act accessed 1 May 2023 and European Commission, 'The Digital

regime set by the ECD has proven to be insufficient to address the complex and dynamic nature of online services and the risks and problems they pose to users and society.

One of these problems is the lack of transparency and accountability of online platforms regarding their content moderation policies and practices, and their cooperation with public authorities. The lack of harmonized rules across the EU on the obligations and responsibilities of online platforms, resulting in legal fragmentation[96], and regulatory uncertainty along with the lack of effective enforcement and cooperation mechanisms created the need for a new legislation to accommodate these risks. As a result of this need, the DSA has emerged, and it aims to address these problems by introducing a new set of rules that apply across the whole EU and that are proportionate to the role, size, and impact of different online service providers.[97]

The DSA differs from the ECD in several aspects. First, it covers a broader range of online intermediaries, including not only hosting providers but also online platforms.[98] Second, it imposes more specific obligations on online intermediaries[99], such as due diligence, transparency, and accountability measures. Third, it introduces a new governance system for the enforcement of the rules, involving national authorities[100], a European board[101] and a digital service coordinator[102].

The European Commission defines "intermediary service providers" as those who provide network infrastructure i.e. Internet access providers.[103] The DSA covers four types of providers as follows: intermediary services[104], hosting services, online platforms[105], and very

---

Services Act Package' (2023) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed 12 August 2023

[96] Ilaria Buri and Joris van Hoboken, 'The Digital Services Act (DSA) proposal: a critical overview', Digital Services Act (DSA) Observatory, Institute for Information Law (IViR), University of Amsterdam, Discussion Paper (28 October 2021) https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf accessed 15 May 2023

[97] European Commission, 'Questions and Answers: Digital Services Act' (2022) <https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348> accessed on 20 November 2022

[98] Digital Services Act, Article 3

[99] Digital Services Act, Articles 5-19

[100] Digital Services Act Article 49

[101] Digital Services Act Articles 61-63

[102] Digital Services Act Article 50

[103] Aina Turillazzia, Mariarosaria Taddeo, Luciano Floridia, Federico Casolari, 'The digital services act: an analysis of its ethical, legal, and social implications', LAW, INNOVATION AND TECHNOLOGY, 2023, VOL. 15, NO. 1, 83–106 (10 Oct 2022) https://doi.org/10.1080/17579961.2023.2184136 accessed 17 May 2023

[104] Digital Services Act, Article 3(g)

[105] 'Considering the particular characteristics of the services concerned and the corresponding need to make the providers thereof subject to certain specific obligations, it is necessary to distinguish, within the broader category of providers of hosting services as defined in this Regulation, the subcategory of online platforms.' Digital Services Act, Recital 13

large online platforms.[106] Recital 15 of the DSA states that this determination of qualification is based on specific activities rather than the entire service provider.[107] This implies that a service provider can be classified as an online platform for certain activities while being considered a "mere conduit" provider for others.[108]

In Chapter II of the DSA, liabilities of the providers of intermediary services are laid down. Under article 4 for "Mere Conduit", the DSA states that service providers are not held liable for the information transmitted or accessed if they meet certain conditions: (a) they do not initiate the transmission, (b) they do not select the receiver of the transmission, and (c) they do not select or modify the information transmitted. Under article 5 for "Caching", service providers are not liable for the temporary storage of user-provided information during its transmission if the conditions of (a) not modifying the information, (b) complying with access and updating rules, (c) not interfering with technology use, and (d) promptly removing or disabling access upon notification of removal or legal orders are met. Last but not least, article 6 states that service providers are not liable for (a) the temporary storage of user-provided information in an information society service, as long as they do not have knowledge of illegal activity or content and (b) if they become aware of such content, they must promptly remove or disable access to it. However, the exemption does not apply if the user is under the control or authority of the service provider. Additionally, in the third paragraph of the article it is stated that the exemption does not cover consumer protection laws for online platforms that may mislead consumers into thinking the platform itself or its affiliated users are the providers of the information or products/services being transacted. After giving the criteria for liability exceptions, the DSA imposes due-diligence obligations applying to all providers of intermediary services[109] in Chapter III and also provides increased obligations for hosting providers and online platforms.[110] Regarding online platforms, the DSA imposes the obligations for managing and reporting complaints to supervisory authorities.[111] It also introduces alternative ways to resolve disputes outside of

[106] Digital Services Act, Recital 41, and European Commission, 'The Digital Services Act: ensuring a safe and accountable online environment: What are the key goals of the Digital Services Act?' https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#what-are-the-key-goals-of-the-digital-services-act accessed 1 May 2023
[107] Digital Services Act, Recital 15
[108] Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021 http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022
[109] *Ibid*
[110] *Ibid*
[111] *Ibid*

court, such as trusted flaggers[112] which will be analyzed in the next chapter, while taking precautions to prevent the misuse of complaints.[113]

In the same chapter, under Section 4, the DSA includes obligations regarding managing risks, providing access to data, ensuring compliance and transparency, and implementing an independent audit for the very large online platforms, in addition to the existing obligations applicable to all online platforms.[114] In Recital 75 of the DSA, it states that considering the significance of very large online platforms, given their extensive user base and their role in promoting public discourse, facilitating economic transactions, and spreading information, and ideas to the public, as well as shaping how users access and share online information, it becomes essential to enforce specific obligations on the providers of such platforms.[115] It also emphasizes that due to their crucial function in locating and making online information accessible, these obligations, to the extent that they are applicable, should also be extended to the providers of very large online search engines. This approach results in imposing additional risk management and transparency obligations for them, and it is important to implement appropriate, proportionate, and effective measures to mitigate the risks, which includes adjusting content moderation and recommendation systems, restricting advertising, enhancing internal oversight, modifying collaboration with trusted flaggers, and establishing cooperation with other platforms through codes of conduct and crisis protocols. [116]

The system for these obligations has a layered structure, a pyramid with horizontal layers that govern different types of services.[117] The first and largest layer applies to all intermediary services and includes the general obligations.[118] The next layer pertains to hosting services, which involve storing user-provided information, and has additional obligations.[119] Online platforms form a distinct subcategory within hosting services, as they not only store but also disseminate[120] user information to the public[121], resulting in another

[112] Digital Services Act, Article 22
[113] Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021 http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022
[114] Digital Services Act, Articles 64 to 80
[115] Digital Services Act, Recital 75
[116] Digital Services Act, Article 35
[117] Folkert Wilman, 'The Digital Services Act (DSA) - An Overview' (16 Dec 2022), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304586 accessed 18 May 2023
[118] *Ibid*
[119] *Ibid*
[120] Digital Services Act, Article 3
[121]Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021 http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022

layer of obligations. The top layer focuses on "very large online platforms" and is determined by the size of the platform, specifically the number of monthly active users.[122] The DSA sets the threshold between "ordinary" online platforms and "very large" platforms at approximately 45 million users, roughly equivalent to 10% of the EU population.[123] The DSA imposes cumulative obligations, meaning that certain services may be subject to requirements from multiple layers of the pyramid.[124] For instance, providers of very large online platforms must adhere to the specific rules for such platforms as well as the regulations for online platforms, hosting services, and intermediary services in general, because very large online platforms fall into all of these categories.[125] On the other hand, services like internet access, classified as mere conduit services, are only accountable for the basic obligations at the bottom layer of the pyramid since they do not fall within the scope of the higher layers.[126]

This creation of four layers of obligations for online service providers is an advancement in regulating digital platforms within EU law.[127] Companies are classified into different types based on some criteria related to size, allowing supervising authorities to target the gatekeeper's business model directly[128], which increases the control of authorities in regulating them. The obligations which the DSA imposes for different online players are proportionate to their role, size and impact in the online ecosystem.[129] Micro and small businesses will be required to fulfill obligations that are proportionate for their capacity and scale, while still maintaining their accountability.[130] In contrast to the ECD, the establishment of the service provider, whether within the EU or a third country, is not a determining factor under the DSA. The new regulations require all online service providers operating in the

[122] Digital Services Act, Recital 77
[123] Digital Services Act, Recital 76
[124] Folkert Wilman, 'The Digital Services Act (DSA) - An Overview' (16 Dec 2022), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304586 accessed 18 May 2023
[125] *Ibid*
[126] *Ibid*
[127] Aina Turillazzia, Mariarosaria Taddeo, Luciano Floridia, Federico Casolari, 'The digital services act: an analysis of its ethical, legal, and social implications', LAW, INNOVATION AND TECHNOLOGY, 2023, VOL. 15, NO. 1, 83–106 (10 Oct 2022) https://doi.org/10.1080/17579961.2023.2184136 accessed 17 May 2023
[128] *Ibid*
[129] European Commission, 'The Digital Services Act: ensuring a safe and accountable online environment: What are the key goals of the Digital Services Act?' https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#what-are-the-key-goals-of-the-digital-services-act accessed 1 May 2023
[130] *Ibid*

single market, regardless of their location in or outside the EU, to adhere to the rules[131], which will increase the applicability of the DSA more.

2.4. Conclusions

The DSA represents a significant legal shift in content moderation in the digital age by reshaping the role of platform operators. While the ECD established the foundation for digital regulation, the DSA addresses the challenges posed by the transformed online environment and aims to create a safer digital space where users' fundamental rights are protected.

Under the ECD, online platforms were granted liability exemptions as long as they promptly removed or disabled access to illegal content upon obtaining knowledge of it. The directive also prohibited general monitoring obligations, striking a balance between the interests of intermediary services, the public interest in removing illegal content, and the respect for conflicting fundamental rights. However, the ECD lacked clear definitions of illegal content and harmonized rules, resulting in legal ambiguity and disparities across Member States.

The DSA expands the scope of intermediaries, covering hosting providers and online platforms, and introduces specific obligations such as due diligence, transparency, and accountability measures. It establishes a layered structure of obligations, with different requirements for various types of intermediaries based on their activities and size. Very large online platforms face additional risk management and transparency obligations due to their significant impact on public discourse, economic transactions, and information sharing. By broadening the range of intermediaries and imposing proportionate obligations, the DSA aims to enhance regulatory control over digital platforms within the EU. It applies to all online service providers operating in the single market, regardless of their location, thereby increasing its applicability and effectiveness in regulating the digital landscape. Overall, the DSA represents a comprehensive and updated framework that addresses the challenges and complexities of content moderation in the digital age. It strives to strike a balance between protecting users' rights, combating illegal content, and fostering a fair and transparent online environment.

Building upon the exploration of the DSA in this chapter, the following chapter will delve into the topic of trusted flagger systems for content moderation. This chapter aims to

[131] *Ibid,* and Folkert Wilman, 'The Digital Services Act (DSA) - An Overview' (16 Dec 2022), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304586 accessed 18 May 2023

examine the concerns, advantages, and disadvantages associated with these systems. It will delve into the workings of trusted flagger systems, highlighting their role in identifying and addressing potentially problematic content. Additionally, it will analyze the ethical and legal challenges posed by trusted flagger systems, ultimately providing a comprehensive understanding of their impact on content moderation in the digital landscape.

CHAPTER III - MAIN ACTOR ON THE STAGE: TRUSTED FLAGGERS

3.1. Chapter Introduction

The development of user-generated content in the rapidly changing digital environment has created both previously unprecedented opportunities for self-expression and the spread of illegal content. Within the framework of the DSA, the concept of flaggers has developed as a potential solution to address these issues. By identifying and notifying online platforms of illegal content, trusted flaggers play a crucial part in content moderation.

This chapter will conduct a thorough analysis to shed light on the complexities of the trusted flaggers. In section 3.2, a legal analysis on the concept of trusted flaggers under the DSA will be conducted. After that, the advantages of using trusted flaggers to moderate online content will be analyzed under section 3.2.1., while the possible drawbacks are being explained under section 3.2.2. Section 3.3 will delve into the legal challenges and concerns surrounding them and section 3.4. will conclude the exploration of the trusted flaggers system and its implications for online content moderation as well as the fundamental right of freedom of expression.

3.2. What exactly Trusted Flaggers are according to the DSA

The rapid advancements in technology and digitalization have transformed users from passive recipients of content to active content creators on various platforms, utilizing text, images, videos, and audio to express their views and promote themselves. However, this shift has also brought to light the negative aspects of digital platforms, such as the easy access to illegal or copyright-infringing content, and the spread of hate speech and terrorist propaganda.[132] These challenges have created difficulties in regulatory practice and law enforcement, particularly when it comes to addressing illegal online content that crosses borders.[133]

To tackle these issues, the concept of trusted flaggers has emerged as a potential solution in content moderation. Trusted flaggers are entities that are granted a special status under the DSA by the Digital Service Coordinator of the Member State they are established in to assist in the notice and action mechanism for tackling illegal content online.[134] This

[132] Mark D. Cole, Christina Etteldorf and Carsten Ullrich, 'Updating the Rules for Online Content Dissemination: Legislative Options of the European Union and the Digital Services Act Proposal' 19 May 2021, https://www.researchgate.net/publication/351667933_Updating_the_Rules_for_Online_Content_Dissemination_Legislative_Options_of_the_European_Union_and_the_Digital_Services_Act_Proposal accessed 7 June 2023

[133] *Ibid*

[134] Digital Services Act, Recital 61

status grants them certain privileges and responsibilities within the notice and action mechanism.

It is important to define "illegal content" to understand what trusted flaggers will exactly detect and disseminate. Under Recital 12 of the DSA, the concept of illegal content is defined broadly to encompass information that reflects existing offline rules. It includes information related to illegal content, products, services, and activities. The term illegal content refers to information, regardless of its form, that is either inherently illegal under applicable law (e.g., illegal hate speech, terrorist content, unlawful discriminatory content) or is rendered illegal because it relates to illegal activities.[135] Illegal content encompasses various types of prohibited material, such as the dissemination of images depicting child sexual abuse, the unauthorized and unlawful sharing of private images, engaging in online stalking, the sale of counterfeit or non-compliant products, offering products or services that violate consumer protection laws, the unauthorized use of copyrighted material, and engaging in illegal activities related to the provision of accommodation services or the sale of live animals.[136]

After defining what illegal content is, the role of trusted flaggers is outlined in the DSA to ensure the timely and effective processing of notices related to illegal content.[137] Providers of hosting services, including online platforms, are required to have easily accessible and user-friendly notice and action mechanisms. [138] The primary function of trusted flaggers is to submit notices to hosting service providers regarding specific items of information that they deem to be illegal content.[139] These notices must be sufficiently precise and adequately substantiated.[140] The DSA specifies that the notices should contain an explanation of why the flagged content is considered illegal, a clear indication of the location of the content, and the identity of the entity submitting the notice, except in cases related to offenses specified in Directive 2011/93/EU[141].[142] These mechanisms allow individuals or

[135] Digital Services Act, Recital 12
[136] Sara Ataei, Jens Van Lathem and Roeland Moeyersons, 'The Practical Implications of the Regulations on Digital Services and Markets' 08 May 2023, https://seeds.law/en/news-insights/the-practical-implications-of-the-regulations-on-digital-services-and-markets/ accessed 10 June 2023
[137] Digital Services Act, Recital 52
[138] Digital Services Act, Recital 50
[139] Digital Services Act, Recital 53
[140] *Ibid*
[141] Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA *OJ L 335, 17.12.2011, p. 1–14.*
[142] Digital Services Act, Recital 53

entities to notify the hosting service provider of specific items of information that they believe to be illegal content.

When trusted flaggers submit notices, they are expected to have particular expertise and competence in detecting, identifying, and notifying illegal content.[143] They must carry out their activities diligently, accurately, and objectively.[144] Trusted flaggers can be entities of various types, including public entities such as internet referral units of law enforcement authorities or non-governmental organizations and private or semi-public bodies dedicated to combating specific types of illegal content, such as organizations involved in reporting child sexual abuse material or illegal racist and xenophobic expressions online.[145]

Trusted flaggers status is regulated under Article 22 of the DSA. This status allows them to have their notices treated with priority by online platform providers, who are required to process and decide upon these notices without undue delay.[146] This means that trusted flaggers' reports should be handled in a timely manner, taking into account the type of illegal content being reported and the urgency of action required. For instance, immediate action should be taken by the hosting service providers when notified about allegedly illegal content that poses a threat to life or safety.[147]

Furthermore, trusted flaggers are expected to publish easily comprehensible and detailed reports on the notices they submit.[148] These reports should provide information such as the number of notices categorized by the hosting service providers, the type of content reported, and the action taken by the providers.[149] The processing of notices submitted by trusted flaggers is expected to be less burdensome and faster due to their demonstrated expertise and competence.[150] However, the time taken to process notices may still vary depending on factors such as the type of illegal content and the quality of the notices.[151] It is important to note that the trusted flagger status should not prevent online platform providers from giving similar treatment to notices submitted by entities or individuals without this status.[152] Providers can also cooperate with other entities in accordance with the applicable

[143] Digital Services Act, Recital 61
[144] *Ibid*
[145] *Ibid*
[146] Digital Services Act, Article 22
[147] Digital Services Act, Recital 52
[148] Digital Services Act, Recital 62
[149] *Ibid*
[150] *Ibid*
[151] *Ibid*
[152] *Ibid*

law.[153] As the DSA recognizes that different types of illegal content may require different processing timelines, and providers should adapt their measures accordingly.[154] Additionally, providers of very large online platforms are encouraged to take appropriate measures, including cooperating with trusted flaggers, to ensure the efficient removal of harmful content, especially content that constitutes cyber violence or non-consensual sharing of intimate or manipulated material.[155]

To prevent misuse of the notice and action mechanism, safeguards should be in place.[156] Misuse can involve frequent submission of manifestly unfounded notices or complaints.[157] The DSA also emphasizes the need for effective safeguards that respect the rights and legitimate interests of all parties, including the fundamental right of freedom of expression.[158] Notices or complaints should be considered manifestly unfounded when it is evident, without substantive analysis, that the content reported is illegal or the notices or complaints lack a valid basis.[159]

3.2.1. The advantages of using the Trusted Flaggers System

As explained above, trusted flaggers are not going to be chosen by the platforms anymore, but they will be appointed by the Digital Services Coordinator of the relevant EU member state.[160] This status of being a trusted flagger is going to be recognized by all online service providers in the scope of the DSA.[161] One possible advantage of having trusted flaggers appointed by the Digital Services Coordinator of the relevant EU member state is that it could enhance the consistency and effectiveness of the enforcement of the rules laid down in the DSA. By having trusted flaggers recognized by all online service providers in the scope of the DSA, it could ensure a more harmonized approach to tackling illegal content across the EU.

Another advantage of using trusted flaggers is their expertise[162] in detecting illegal content. Trusted flaggers have expertise in detecting illegal content which can help online

---

[153] *Ibid*
[154] *Ibid*
[155] Digital Services Act, Recital 87
[156] Digital Services Act, Recital 63
[157] *Ibid*
[158] *Ibid*
[159] *Ibid*
[160] Tremau, 'New Role of Trusted Flaggers In The EU', 25 May 2022, https://tremau.com/digital-services-act-trusted-flagger-organisations accessed 10 June 2023
[161] *Ibid*
[162] Digital Services Act, Article 22/1

platforms to identify and remove illegal content more efficiently[163], by doing it more quickly and reliably.[164] It can improve the quality and accuracy of flagging, by relying on the expertise and competence[165] of trusted flaggers in specific areas, such as terrorism, child abuse, hate speech, or intellectual property rights.[166] Under Article 22/1 of the DSA, because it is projected that the notices from trusted flaggers are given priority and are processed and decided upon without undue delay, it can also help online platforms to take action more quickly ("expeditiously"). Platforms like YouTube, Twitter and TikTok explicitly mentioned and used trusted flaggers in their content moderation practices and they gave these notifiers privileges voluntarily.[167] This prioritization aims increasing actionability[168] and the DSA has the same goal which aims to take action more quickly. By adopting a coherent, centralized and more harmonized approach, the cooperation between online platforms and competent authorities will be systemically enhanced, as well as the cross-border cooperation between the Member States.[169]

Overall, trusted flaggers play an important role in content moderation by reporting illegal content to online platforms for review. Since they have expertise in detecting content, online platforms will be able to identify and remove illegal content more efficiently and because their reports will be reviewed with priority, more quickly.

3.2.2. The Potential Drawbacks of the Trusted Flaggers

Trusted flaggers can be "trusted" provided that they are separate actors who operate independently from online platforms, all kinds of commercial entities, and law enforcement agencies.[170] Their mission should revolve around serving the collective interests of the public

---

[163] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

[164] European Commission, 'Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Tackling Illegal Content Online Towards and Enhanced Responsibility of Online Platforms', COM/2017/0555 final

[165] Digital Services Act, Article 22/1. a

[166] Kerstin Bäcker, 'The trusted flaggers in the Digital Services Act', 23 May 2022, https://www.lausen.com/en/the-trusted-flaggers-in-the-digital-services-act/ accessed 11 June 2023

[167] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

[168] YouTube, 'About the YouTube Priority Flagger program' https://support.google.com/youtube/answer/7554338?hl=en#zippy= accessed 11 June 2023

[169] European Commission, 'Communication from The Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Tackling Illegal Content Online Towards and Enhanced Responsibility of Online Platforms', 28 September 2017

[170] EDRi, 'The Digital Services Act, The EDRi guide to 2,297 amendment proposals' October 2021, https://edri.org/wp-content/uploads/2021/10/EDRi-policy-paper-Digital-Services-Act-Nov-2021.pdf accessed 11 June 2023

and safeguarding fundamental rights.[171] The criteria for awarding the status of trusted flaggers are defined in Article 22 of the DSA, however they can be more clear to avoid leading to a lack of transparency in the reporting process.[172]

The risk of over-blocking the online content is another drawback of using trusted flaggers. There is a possibility that flaggers may be prone to over-blocking, meaning they remove content excessively, and the mechanisms to hold them accountable for their actions might not be sufficient.[173] This exacerbates the concerns regarding the transparency and accountability of content moderation practices.[174] This lack of accountability not only undermines the transparency and fairness of content moderation practices but also raises doubts about the reliability and effectiveness of trusted flaggers in striking a balance between removing illegal content and protecting freedom of expression.

Trusted flagging tends to serve the interests of established public and private powers like law enforcement and intellectual property rights holders.[175] These are examples of parties which have major influences and might demand recognition from platforms in line with their own political and economic interests.[176] This will potentially raise questions about the "independence" of trusted flaggers, even the DSA aims to provide a regulatory oversight. Under Article 22 paragraph 6 of the DSA, if an online platform provider becomes aware that a trusted flagger has submitted a significant number of notices that are imprecise, inaccurate, or lack sufficient substantiation, they are required to inform the Digital Services Coordinator who granted trusted flagger status to that entity. The platform provider must provide relevant information, explanations, and supporting documents. Upon receiving this information, the Digital Services Coordinator can initiate an investigation *if there are legitimate reasons to do so*, leading to the suspension of trusted flagger status during the investigation period.[177] The absence of clear definition of these "legitimate reasons" for the Digital Services Coordinator to suspend the trusted flagger status, raises concerns about the accountability of trusted flaggers. To ensure a more accountable system, it is imperative to establish clearer safeguards and criteria that outline the specific circumstances under which trusted flagger status can be

---

[171] *Ibid*
[172] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022
[173] *Ibid*
[174] *Ibid*
[175] *Ibid*
[176] *Ibid*
[177] Digital Services Act, Article 22/6

suspended. These safeguards would enhance transparency, fairness, and the overall integrity of the trusted flagging process.

Overall, these drawbacks raise concerns about the effectiveness, fairness, and potential abuse of trusted flagging systems, necessitating the implementation of effective safeguards and regulatory measures.

3.3. The Legal Challenges and Concerns Surrounding the Use of Trusted Flaggers

The legal challenges associated with the use of trusted flaggers in online content moderation are multidimensional. As much as the trusted flaggers system is intended to improve the efficiency and accuracy of the content moderation, as well as to enhance the accountability and transparency of online platforms[178], it also raises several legal challenges, especially in relation to freedom of expression and other human rights. These challenges revolve around issues like enforcement overreach, the "illegality" status of the flagged content, concerns about over-blocking and lack of accountability, and the lack of transparency and accountability in trusted flagging procedure.

One of the primary concerns is the imposition of mandatory usage of trusted flaggers and the recognition of their notices as constituting "actual knowledge" of illegal content.[179] By making their usage mandatory, trusted flaggers would assume quasi-judicial functions, potentially impeding the effectiveness of this approach.[180] Furthermore, considering their notices as "actual knowledge" could place content platforms at risk of legal liability for not acting upon such information, even if it is later determined to be incorrect or unsubstantiated. This raises questions about the legal implications and fairness of relying solely on trusted flaggers for determining the legality of content.

Firstly, giving trusted flaggers the priorities explained above raises the question of enforcement overreach.[181] The DSA enables law enforcement agencies, such as the European

---

[178] IMCO Committee, 'Online Platforms' Moderation of Illegal Content', (2020), https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf accessed 21 November 2022

[179] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[180] *Ibid*

[181] Konstantinos Komaitis, Katitza Rodriguez and Christoph Schmon, 'Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act' Electronic Frontier Foundation, 18 February 2022 https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act accessed 10 June 2023

Union Agency for Law Enforcement Cooperation ('Europol') as it is explicitly mentioned[182] in the regulation, to apply to be a trusted flagger. This means that they will have the privilege of requesting the removal of content, which can be argued that this will result in increasing the power of a law enforcement agency by the hand of trusted flagger system.[183] On 3 January 2022, the European Data Protection Supervisor ('EDPS') ordered Europol to delete the data of individuals who have no confirmed link to a criminal activity, which poses a clear risk to fundamental rights.[184] The fact that Europol keeping the data for longer than necessary in that case raises the question of surveillance of European citizens.[185] Therefore, it can be assumed that such empowerment of a law enforcement agency will also infirm the credibility of trusted flaggers in the case of content moderation, and consequently will raise the question of "How much independency?". Under Recital 61 of the DSA, trusted flaggers can be public in nature, non-governmental organizations and private or semi-public bodies. Since each Member State has its own notion of "illegal", privatization of the enforcement might also cause an unpredictable environment and could stimulate forum shopping[186], which will result in the opposite of harmonized application and enforcement of the regulation and lead to potential biases and imbalances in content moderation.

The risk of over-blocking of content is another legal issue which is related to freedom of expression concerns. One argument is that trusted flagging exacerbates existing issues related to transparency, accountability, and contestability in content moderation practices.[187] When approached from a governance perspective, highlighting the technocratic nature of the approach that emphasizes "expertise" without considering other forms of representativeness,

---

[182] Digital Services Act, Recital 61

[183] Konstantinos Komaitis, Katitza Rodriguez and Christoph Schmon, 'Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act' Electronic Frontier Foundation 18 February 2022 https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act accessed 10 June 2023

[184] European Data Protection Supervisor, 'EDPS orders Europol to erase data concerning individuals with no established link to a criminal activity' 10 Jan 2022, https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-orders-europol-erase-data-concerning_en#:~:text=On%203%20January%202022%2C%20the,EDPS'%20inquiry%20launched%20in%202019. accessed 09 June 2023

[185] Konstantinos Komaitis, Katitza Rodriguez and Christoph Schmon, 'Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act' Electronic Frontier Foundation 18 February 2022 https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act accessed 10 June 2023

[186] *Ibid*

[187] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

such as cultural, political, and socioeconomic factors.[188] This might result in trusted flaggers prioritizing removal over protecting freedom of expression and potentially leading to biases.[189] Another challenge stems from the misaligned incentives between platforms and trusted flaggers.[190] Institutionalized trusted flagging programs may exacerbate trends of over-removal of content, as platforms may lack adequate incentives to combat such practices.[191] In fact, platforms might attempt to justify their removal actions by shifting responsibility onto the trusted flaggers.[192] Additionally, when an actor's role as a trusted flagger aligns with its economic interests, such as intellectual property rights-holders, or bypasses constitutional safeguards for government action, concerns arise regarding potential 'privatized censorship' and the erosion of due process.[193]

In conclusion, the legal challenges associated with the use of trusted flaggers in online content moderation encompass a wide range of legal concerns. Legal challenges surrounding the use of trusted flaggers involve issues related to freedom of expression, enforcement overreach, potential biases, and erosion of due process. Balancing effective content moderation with user rights and freedom of expression necessitates strong safeguards and regulatory measures. Addressing these challenges requires striking a balance between effective content moderation, freedom of expression, fairness, and the preservation of user rights in the digital realm.

3.4. Conclusions

Trusted flaggers have emerged as an instrument within the framework of the DSA to tackle the challenges posed by illegal content dissemination in the digital landscape. The DSA recognizes their expertise in detecting and notifying illegal content to online platforms, granting them a special status which enables them to submit notices with priority for review and action. This approach aims to improve the efficiency and accuracy of content moderation while fostering a safer online environment.

However, the implementation of trusted flaggers also raises certain concerns and legal challenges. The potential for over-blocking and lack of accountability in content removal decisions, as well as the potential for enforcement overreach and biases, must be carefully

[188] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed on 25 November 2022
[189] *Ibid*
[190] *Ibid*
[191] *Ibid*
[192] *Ibid*
[193] *Ibid*

addressed to ensure the protection of fundamental rights, including freedom of expression. Striking a balance between the effectiveness of content moderation and the preservation of user rights remains crucial in establishing a robust and trustworthy trusted flagger system.

Building upon the analysis of trusted flaggers, advantages and possible drawbacks of them, and the legal challenges and concerns around using them in this chapter, the following chapter will focus on the implementation of the trusted flagger system for content moderation, specifically exploring its effectiveness, proportionality, and offering recommendations for its application. The chapter will discuss different ways of implementing trusted flagger systems, including various models and criteria that can be considered. By providing insights and recommendations, the chapter seeks to contribute to the optimization of trusted flagger system implementation, fostering a safer and more inclusive online environment while respecting user's fundamental rights, especially the freedom of expression.

CHAPTER IV - HOW TO IMPLEMENT THE TRUSTED FLAGGERS SYSTEM FOR CONTENT MODERATION

4.1 Chapter Introduction

To understand how trusted flaggers operate and their place in the content moderation process, first the different models and practices of trusted flagging, depending on the type and level of involvement of public and private actors in setting and enforcing the rules and standards for content moderation will be introduced in section 4. 2. Second, the challenges and opportunities of balancing content moderation and freedom of expression will be discussed under the section 4.3. The principles and criteria that apply to any restriction on freedom of expression, and how they can be respected and implemented in content moderation practices will be explained. Finally, in 4.4, recommendations to ensure that trusted flaggers are used effectively and accountably under the DSA will be suggested.

4.2. Different Practices of Trusted Flagging

There are different models for trusted flagging, depending on the type and level of involvement of public and private actors in setting and enforcing the rules and standards for content moderation. These models can be categorized into three main types as follows: (i) self-regulatory model,[194] (ii) co-regulatory model,[195] and (iii) regulatory model[196].

4.2.1 Self-regulatory Model

In this model, online platforms and trusted flaggers voluntarily enter into partnerships to cooperate and coordinate on content moderation issues without any formal or legal obligation or oversight by public authorities.[197]

Two examples for this model can be YouTube's Trusted Flagger Program and Facebook's Third-Party Fact Checking Program.[198] YouTube's Trusted Flagger Program is a voluntary partnership between YouTube and selected organizations which are experts in

---

[194] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

[195] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[196] German Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, 1 September 2017

[197] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

[198] *Ibid*

identifying content that violates YouTube's Community Guidelines.[199] These organizations include NGOs and government agencies that work on issues like child safety, human rights, terrorism or hate speech.[200] These flaggers have access to a special reporting tool that allows them to flag multiple videos at once, as well as a support team that reviews their flags with priority. According to YouTube, trusted flaggers have a high accuracy rate of over 90% in flagging content violating its policies.[201] Facebook's Third-Party Fact-Checking Program is another example for this model which is a collaboration between Facebook and independent fact-checkers which are certified by the International Fact Checking Network ('IFCN').[202] Facebook says that because a private company like Meta should not be deciding on what is right or wrong, they choose to work with "non-partisan global fact-checking partners who independently work".[203]

Both examples show that this is a model in which online platforms and trusted flaggers voluntarily cooperate and coordinate on content moderation, without any formal or legal obligation or oversight by public authorities. The advantage of this model can be allowing for flexibility[204], efficiency, and innovation in content moderation, as well as for the recognition of the competence of trusted flaggers. However, it might lack transparency and accountability in content moderation which will raises concerns about the legitimacy and representativeness[205] of trusted flaggers. Therefore, this model may not be sufficient or effective in addressing all types of illegal content, or in balancing the rights and interests of stakeholders involved in content moderation.

---

[199] Youtube Help, 'About the YouTube Priority Flagger Program' https://support.google.com/youtube/answer/7554338?hl=en accessed 11 June 2023

[200] *Ibid*

[201] Jen Carter, Growing Our Trusted Flagger Program Into YouTube Heroes, 22 September 2016, https://blog.youtube/news-and-events/growing-our-trusted-flagger-program/ accessed 29 June 2023

[202] Meta, 'How Meta's third-party fact-checking program works' 1 June 2021 https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works accessed 29 June 2023

[203] *Ibid*

[204] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 *p.9* accessed 12 June 2023

[205] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf *p.15* accessed 25 November; and Sebastian Felix Schwemer, 'Trusted Notifiers and the Privatization of Online Enforcement' *Computer Law & Security Review, Volume 35, Issue 6,* (2019), https://doi.org/10.1016/j.clsr.2019.105339 accessed 28 November 2022

4.2.2 Co-regulatory Model

In this model, online platforms and trusted flaggers cooperate and coordinate on content moderation issues, but also adhere to certain rules and standards set or endorsed by public authorities to set and enforce rules and standards for online content.[206] The co-regulatory model can take different forms, such as codes of conduct, self-regulation schemes, or independent oversight bodies, but it usually involves a degree of consultation, coordination and cooperation between the public and the private actors involved in content moderation.[207] Under Recital 14 of the EU Audiovisual Media Services ('AVMS') Directive, co- regulation is defined as *"a legal link between self-regulation and the national legislator in accordance with the legal traditions of the Member States"* and the stakeholders and the government or the national regulatory authorities share the regulatory role in this model.[208]

One example for this model is the EU Code of Conduct on Countering Illegal Hate Speech Online[209] launched in 2016 as a voluntary agreement between the European Commission and Facebook, Microsoft, Twitter and YouTube.[210] The Code of Conduct aims to ensure that illegal hate speech is removed or disabled within 24 hours of notification by trusted flaggers ("reporters").[211] It also sets standards for transparency, feedback, counter-narratives, and education on hate speech.[212] This example illustrates how the co-regulatory model works by creating a common framework for online platforms to remove illegal content online, while also allowing assessment of their content moderation practices by the Commission.

The co-regulatory model also offers several benefits and challenges. On the one hand, it can enhance the legitimacy, accountability, and diversity of content moderation decisions

---

[206] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[207] *Ibid, p.11-12*

[208] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU 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) in view of changing market realities, *OJ L 303, 28.11.2018, p. 69–92*

[209] The classification of this model is intricate in literature. According to some opinions in the literature, this model is classified under self-regulatory model as well, since both models are generally overlaps in some regards. See: Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf

[210] EU Code of Conduct on countering illegal hate speech online – European Commission, May 2016, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en accessed 2 July 2023

[211] Code of Conduct on Countering Illegal Hate Speech Online, p.3, accessed 2 July 2023

[212] *Ibid, p.3*

by including a wider range of stakeholders and perspectives.[213] On the other hand, it can also pose some challenges in balancing the autonomy and responsibility of online platforms with the public interest and legal certainty of content moderation.

4.2.3 Regulatory Model

The regulatory model involves legal obligations for online platforms to detect and remove illegal content according to the rules set by the public authorities.[214] The early attempt for this model for regulating the online content by the hands of trusted flaggers is the German Netzwekdurchsetzungsgesetz ('NetzDG')[215], which entered into force on January 1 2018[216]. It applies to large online platforms that have more than 2 million users located in Germany and requires them to enable their users to submit notices about illegal content.[217] Once they receive a notice, platforms are required to check if the reported content is illegal.[218] They must remove the content if it is clearly illegal *("manifestly unlawful")* within 24 hours and other illegal content must be removed within 7 days. Platforms that fail to follow these rules may face fines up to €50 million.[219]

One of the advantages of this model is that it has a clear legal framework for content moderation in the national level, based on the existing criminal code of Germany[220] in the example. It also enhances transparency and accountability of online platforms by requiring them to publish reports on their content moderation practices and also to set up complaint mechanisms for users.[221] However, it also might undermine freedom of expression and creating a chilling effect on online speech, as platforms may over-remove for the sake of

---

[213] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[214] Rachel Griffin, 'New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG' SSRN Electronic Journal 2021 https://sciencespo.hal.science/hal-03586791/document accessed 1 July 2023

[215] Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers' Yale Law School https://yjolt.org/sites/default/files/0_-_appelman_leerssen_-_on_trusted_flaggers.pdf accessed 25 November 2022

[216] Heidi Tworek & Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 2019, https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf accessed 2 July 2023

[217] *Ibid*

[218] *Ibid*

[219] *Ibid*

[220] Stefan Theil, 'The German NetzDG: A Risk Worth Taking?' 2018 https://verfassungsblog.de/the-german-netzdg-a-risk-worth-taking/ accessed 2 July 2023

[221] *Ibid*

being compliant with the law to avoid fines.[222] For instance, NetzDG requires some associated data to be stored for future investigations, which risks inadvertent results for privacy and other human rights.[223] Another shortcoming of this model is that it transfers the responsibility of deciding what is illegal content from courts and public authorities to private platforms.[224] Last but not least, it might have negative spillover effects on authoritarian regimes, that might use it as a model or justification for stricter laws and regulations on content moderation.[225]

The three models for content moderation have their own advantages and disadvantages in terms of effectiveness, accountability, transparency, and respect for human rights. There is no one-size-fits-all for content moderation, as different types of content might require different approaches and standards. Choosing a model for content moderation should consider the interests of all stakeholders like online platforms, trusted flaggers, public authorities, civil society organizations, and of course users.

4.3 Protecting Freedom of Expression while Moderating Content

When it comes to online content moderation, the two values that need to be balanced are protecting freedom of expression and preventing harm.[226] Online platforms are the primary actors of online speech, and they increasingly rely on automated content moderation by using machine-learning algorithms.[227] This creates additional problems to the already complicated issue at hand like false positives and false negatives.[228] Either completely by the hands of algorithms or with humans in the loop, content moderation requires striking a balance between freedom of expression rights and interests and values of societies.[229]

Freedom of expression is a fundamental human right allowing people to express their opinions, thoughts, and ideas without fear of censorship, punishment, or retaliation. It also

---

[222] Heidi Tworek & Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 2019, https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf accessed 2 July 2023

[223] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[224] *Ibid*

[225] Janosch Delcker, 'Germany's balancing act: Fighting online hate while protecting free speech' 1 October 2020, https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/ accessed 2 July 2023

[226] Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser & Jason Reifler, 'Resolving content moderation dilemmas between free speech and harmful misinformation' 9 November 2022, https://www.pnas.org/doi/epdf/10.1073/pnas.2210666120 accessed 3 July 2023

[227] *Ibid*

[228] *Ibid*

[229] *Ibid*

includes the right to seek, receive, and impart information and ideas through any media and regardless of frontiers.[230] This right is protected under Article 10 European Convention on Human Rights[231] ('ECHR'), which states that "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.". However, because it is not an absolute right, it is subject to certain restrictions caused by law, or for the protection of national security, public order, or public morals.[232] Therefore, to protect the essence of this right, these possible restrictions must comply with the principles of legality, legitimacy, necessity, and proportionality.[233]

In the case of *Delfi AS v. Estonia*, and the subsequent case of *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*, the interplay between the right to freedom of expression and the responsibilities of online platforms in moderating user-generated content comes into focus. These cases provide a real-world context to the complexities of striking the balance between facilitating open discourse and safeguarding against potential harm caused by online content.

In the case of *Delfi AS v. Estonia*[234], the European Court of Human rights ('ECtHR') examined whether an online news portal could be held liable for offensive comments posted by readers below one of its articles. As in the other cases regarding Article 10 of ECHR, The ECtHR applied a three-part test to decide whether Delfi's rights had been violated disproportionately or not.[235] First, because Estonia imposed civil penalties for the defamatory comments, the court found that there was an interference with the outlet's right to freedom of expression.[236] Then the court held that because there was a violation of Estonia's Civil Code Act and Obligations Act by the outlet, the reimbursement of damages was a requirement of law.[237] Third, the court pointed out that the civil penalties imposed on the outlet in the case was a legitimate means of protecting "the reputation and rights of others".[238] Finally, after conducting the balancing test to decide whether this interference of Estonia with the rights of

[230] ECHR, Article 10, and ICCPR, Article 19
[231] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR)
[232] ECHR, Article 10(2)
[233] Gehan Gunatilleke, 'Justifying Limitations on the Freedom of Expression' 1 November 2020, https://link.springer.com/article/10.1007/s12142-020-00608-8, accessed 4 July 2023
[234] Delfi AS v. Estonia, Application no. 64569/09, (2015) EHRR
[235] Global Freedom of Expression, 'Delfi AS v. Estonia', Columbia University https://globalfreedomofexpression.columbia.edu/cases/delfi-as-v-estonia/ accessed 13 August 2023
[236] *Ibid*
[237] *Ibid*
[238] *Ibid*

the outlet was necessary in a democratic society,[239] the ECtHR found that there was no violation of Article 10 of the ECHR, as the portal in question had failed to prevent clearly unlawful comments from being published which amounted to hate speech and incitement to violence.[240]

In contrast, in the case of *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary,* the ECtHR examined whether a self-regulatory body of internet content providers and one of its members, an online news portal, could be held liabl for vulgar comments posted by readers on the portal's website.[241] The domestic courts ruled that the comments were beyond the boundaries of freedom of expression and damaged the right to reputation of the real-estate website and awarded damages.[242] However, the ECtHR ruled that there was a violation of Article 10 of the ECHR and held that Hungary had failed in balancing competing rights (here the right to freedom of expression and right to reputation), when it awarded damages to the real-estate website for the harms to its business reputation.[243] The ECtHR held that the applicants were not publishers or authors of the comments, but merely provided a platform for user-generated content.[244] The court also stressed that the comments did not constitute hate speech or incitement to violence,[245] but rather vulgar criticism of a real estate company. The judgment of the ECtHR here shows that holding web portals accountable for the third-party comments violated the freedom of expression.

The decisions in these cases offer insights into how courts address the challenges regarding protecting freedom of expression while dealing with the possible consequences of uncontrolled online discourse. Content moderation is a form of restriction on freedom of expression, as it involves the removal of certain types of content on online platforms. Therefore, content moderation practices must adhere to the principles of legality, legitimacy, necessity, and proportionality, which apply to any restriction on this right. These judgments emphasize the importance of context, and ensuring whether the restrictions on freedom of expression are necessary, proportionate, and justified.

---

[239] *Ibid*
[240] Delfi AS v. Estonia, Application no. 64569/09, (2015) EHRR
[241] *Ibid*
[242] Global Freedom of Expression, 'Delfi AS v. Estonia', Columbia University https://globalfreedomofexpression.columbia.edu/cases/delfi-as-v-estonia/ accessed 13 August 2023
[243] Global Freedom of Expression, 'Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary', Columbia University https://globalfreedomofexpression.columbia.edu/cases/magyar-tartalomszolgaltatok-egyesulete-index-hu-zrt-v-hungary/ accessed 13 August 2023
[244] Magyar Tartalomszolgáltatók Egyesülete And Index.Hu Zrt v. Hungary, Application no. 22947/13, (2016) EHHR
[245] *Ibid,* prg. 64

When content moderation practices are too strict, it can result in removal of the content that is not illegal, but rather legitimate, or protected by freedom of expression, which creates the problem of over-removal. It can result from the use of automated tools that fail to understand context, nuance, or irony.[246] If the application of the rules and standards are not consistent and clear enough, it might happen as a result of the lack of sufficient safeguards and remedies for users whose content is removed.[247] Without proper and sufficient safeguards, these automated tools might contribute to censorship and biased enforcement of the laws.[248]

It is also important that necessary transparency measures are taken to make clear the reasons behind the removed content, as it would be difficult for individuals to make an appeal.[249] There should be timely, accessible, and also fair appeals processes.[250] If there happens to be a mistake, sufficient transparency measures would prevent the possibility of similar mistakes in the future and this is an essential part of any redress mechanism.[251] It is also essential to recognize potential discrimination in the manner in which complaints regarding various forms of content are addressed.[252] Lack of transparency for content moderation is a major challenge on freedom of expression as it can undermine the trust, legitimacy, and accountability of online platforms, hinder the oversight and evaluation of content moderation decisions, and limit the access to information.[253]

This also brings up the lack of accountability as a major challenge for content moderation and freedom of expression, as it affects the rights and interests of users and other stakeholders who are affected by content moderation decisions and practices. Lack of accountability can violate the due process rights of users whose content is removed, such as

---

[246] Natasha Duarte, Emma Llanso and Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' *Center for Democracy & Technology,* Proceedings of Machine Learning Research 81:1-1 2018 https://proceedings.mlr.press/v81/duarte18a.html accessed 4 July 2023

[247] Thiago Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' Human Rights Law Review 20, 5 December 2020, p.614, https://doi.org/10.1093/hrlr/ngaa032 accessed 29 June 2023

[248] Natasha Duarte, Emma Llanso and Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' *Center for Democracy & Technology,* Proceedings of Machine Learning Research 81:1-1 2018 https://proceedings.mlr.press/v81/duarte18a.html accessed 4 July 2023

[249] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023

[250] *Ibid*

[251] *Ibid*

[252] *Ibid*

[253] Nicolo Zingales, Thales Bertaglia and Catalina Goanta, 'Freedom of expression & Content Moderation on Social Media: The Promises and Challenges of Experimental Accountability' Maastricht University Blog, 10 August 2021, https://www.maastrichtuniversity.nl/blog/2021/08/freedom-expression-content-moderation-social-media-promises-and-challenges-experimental accessed 3 July 2023

the right to be informed, the right to be heard, and the right to appeal.[254] It is directly related to freedom of expression in that it might affect the protection and promotion of this right by online platforms and their content moderation practices. Under Article 13 of the European Convention on Human Rights[255] ('ECHR'), "everyone whose rights and freedoms are violated shall have an effective remedy before a national authority".[256] Human rights must be respected and protected for both those who were the victims of online offenses and those whose freedoms were restricted as a result of content moderation measures.[257]

4.4. How to Implement Trusted Flaggers for Content Moderation - Recommendations

The DSA has many different parts, but at its core it is a law for digital due process that comes with a number of risk management instruments.[258] Universal due process protections are established, private decision-making is encouraged to be transparent, and ongoing risk management by major players is institutionalized.[259] The DSA establishes guidelines for how disputes must be handled and provides individuals with legitimate remedies to challenge the outcomes both internally and externally.[260] It encourages the creative industries to step up their content moderation practices by formalizing the idea of trusted flaggers.[261] However, as it is explained in the previous chapter, serious questions about human rights are raised especially when law enforcement officials are designated as

[254] Council of Europe, 'Guidance Note on Content moderation' 20 September 2021 https://www.coe.int/en/web/freedom-expression/-/guidance-note-on-content-moderation accessed 28 January 2023; and
United Nations Human Rights Office of the High Commissioner, 'Moderating online content: fighting harm or silencing dissent?' 23 July 2021 https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent accessed 12 March 2023
[255] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights, as amended) (ECHR)
[256] ECHR, Article 13
[257] Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18 accessed 12 June 2023
[258] Martin Husovec, 'The DSA as a Creator's Charter?' *Journal of Intellectual Property Law & Practice, 2023, Vol. 18, No. 2* https://academic.oup.com/jiplp/article-pdf/18/2/71/49276302/jpac116.pdf accessed 25 June 2023
[259] *Ibid*
[260] *Ibid*
[261] *Ibid*

trusted flaggers.[262] It can also be possible to reevaluate how judicial oversight can be successfully incorporated into online content moderation according to the DSA.[263]

Implementing trusted flaggers for content moderation while protecting freedom of expression requires a careful balance between combating illegal content and respecting human rights. We need *ex-ante* guidelines for both trusted flaggers and online platforms to provide insights on balancing content moderation and protecting the freedom of expression. These guidelines would also facilitate certainty since that will enable all stakeholders to foresee what kind of interventions they might face.

To guarantee that the DSA's measures are successful and in accordance with human rights standards, these guidelines and delegated acts should be clear.[264] Particularly acknowledged in the DSA's due diligence section are the systemic risks to fundamental rights especially posed by very large online platforms' ('VLOPs') operational and systemic practices.[265] Future delegated acts and guidelines that the Commission has yet to establish will determine how successful the DSA's measures are.[266] To ensure that trusted flaggers are used properly for content moderation that adheres to human rights standards, it can be helpful to provide more explicit instructions and designated tasks. This approach could be more fitting for the co-regulatory model, enabling all stakeholders to be involved in the process with coordination. Especially for online platforms, they would continue to apply their own policies to some extent in accordance with the aforementioned suggestions. This way, it might also cultivate a balance for online platforms, ensuring they fulfill their legal obligations while simultaneously protecting the freedom of expression.

Enhancing the cooperation and coordination among different actors involved in content moderation, such as online platforms, trusted flaggers, competent authorities, Digital Services Coordinators, European Board for Digital Services is also crucial for the DSA to reach its goal of providing a safer online environment while respecting the fundamental rights of the users. The DSA establishes a network of actors with different roles and responsibilities in ensuring a safe and accountable online environment. However, this network may face

---

[262] Iverna McGowan, 'How can we apply human rights due diligence to content moderation? Focus on the EU Digital Services Act – Event Summary' September 9, 2021 https://cdt.org/insights/how-can-we-apply-human-rights-due-diligence-to-content-moderation-focus-on-the-eu-digital-services-act-event-summary/ accessed 7 July 2023

[263] *Ibid*

[264] Eliska Pirkova, 'The Digital Services Act: your guide to the EU's new content moderation rules' 6 July 2022, https://www.accessnow.org/digital-services-act-eu-content-moderation-rules-guide/ accessed 9 July 2023

[265] *Ibid*

[266] *Ibid*

challenges of communication, information-sharing, consistency, or trust among its members. The DSA could facilitate more dialogue and exchange of best practices among these actors, as well as provide more clarity on their respective mandates and competences through European Board for Digital Services.[267] The board will be composed of representatives of national authorities and the Commission[268] to contribute and assisst the consistent application of the DSA.[269] The board could achieve this by issuing guidelines and recommendations, as well as the opinions on different aspects of content moderation such as the criteria for trusted flaggers, and transparency and accountability measures. By doing so, the DSA could achieve this goal of fostering more cooperation and coordination among different actors in the content moderation process.

Last but not least, involving trusted flaggers in the design and implementation of content moderation practices instead of just cooperating with them would also increase the efficiency and reliability of them. Instead of giving trusted flaggers just a reactive role as it is set in the DSA, benefiting from their expertise can improve the proactive measures and the tools used for detecting and removing illegal content. Developing more accurate and nuanced tools and algorithms and collaborating with trusted flaggers to improve the quality and reliability of content moderation process could facilitate protection of right to freedom of expression, and also address the over-blocking problem. By encouraging more communication and collaboration between online platforms and trusted flaggers on how to develop and apply content moderation rules and algorithms, the DSA can be one step closer to its goal of ensuring a safer online environment.

In conclusion, there are many challenges and opportunities of implementing trusted flaggers for content moderation under the DSA. Content moderation practices must comply with the principles of legality, legitimacy, necessity, and proportionality that apply to any restriction of the fundamental right of freedom of expression. We have suggested our recommendations to enhance the effectiveness and accountability of trusted flaggers, which are providing more clear guidelines, fostering the cooperation and coordination among different stakeholders, and involving trusted flaggers in the design and implementation of content moderation practices. By following these recommendations, the DSA can achieve its aim of creating a safer online environment while protecting users and the essence of the right to freedom of expression.

[267] Digital Services Act, Article 61
[268] Digital Services Act, Article 62
[269] Digital Services Act, Article 61

CHAPTER V – CONCLUSION

The present study analyzes the challenges arising from current content moderation practices and the risks they pose to the freedom of expression while examining the responsibilities for the online platform operators to answer the following question:

"How does the implementation of trusted flaggers system under the DSA affect the freedom of expression within the context of online content moderation?"

Based on the insights and findings from the previous chapters, the implementation of the trusted flaggers system for content moderation under the Digital Services Act requires a balanced approach. The comprehensive exploration of this newly introduced actor has revealed a complex landscape. Drawing on the insights and findings from the preceding chapters, it is understood that the trusted flaggers play a pivotal role in addressing the challenges posed by the illegal content online. The total sum of considerations, advantages, and the potential drawbacks regarding the using and implementation of the trusted flaggers has emphasized that a harmonious equilibrium between content moderation practices and the preservation of freedom of expression has utmost importance.

The digital landscape is undergoing a transformative shift with the introduction of the DSA. It represents a significant legal transition from the E-Commerce Directive, and it seeks to address the evolving challenges of content moderation. One of the primary focal points of the DSA is the establishment of a more comprehensive regulatory framework for digital service providers acting as intermediaries. It introduces a layered structure of obligations, tailored to the activities and size of different online intermediaries, thereby aiming to ensure proportionate responsibilities. The shift in intermediary liability places a great emphasis on due diligence, transparency, and accountability, with VLOPs facing additional risk management requirements. The DSA' s approach aims to strike a balance between protecting users' rights, tackling illegal content, and fostering a fair, transparent, and accountable online environment.

Trusted flaggers are the entities granted a special status under the DSA with significant power in the detection and removing of illegal content on online platforms. Especially because of their expertise, using them offers a potentially successful solution for more efficient and reliable identification and dissemination of illegal content, especially in cases of hate speech, terrorist propaganda, child abuse material, and copyright infringement. This increased capacity to address such challenges is a crucial step forward in creating a safer online environment as the DSA aims to achieve.

Taking into consideration the comprehensive analysis of trusted flaggers and content moderation practices presented in this study, it has been found that different practice models have emerged. These practice models could be categorized into three different main types: the self-regulatory model, the co-regulatory model, and the regulatory model. While the DSA itself does not explicitly mention or outline a particular model for trusted flagging, the examination of these models reveals different approaches employed in the complex task of balancing the need of tackling illegal content online while protecting freedom of expression. Therefore, this study suggests that in the application of the DSA, both the advantages and disadvantages of each model and should be considered and an overarching approach that can include the diversity of online platforms should be adopted. The choice and implementation of the content moderation model is ultimately a shared responsibility between online platforms and the public authorities, with the inclusion of civil society and users.

The focal challenge that emerges from this analysis centers on balancing content moderation and freedom of expression. Content moderation is necessary to prevent harm, but it must adhere to principles of legality, legitimacy, necessity, and proportionality. A delicate balance must be maintained to avoid over-removal of legitimate content and the consequent chilling effects on online speech. Transparency and accountability in content moderation practices play a crucial role in upholding users' rights and interests.

To address these challenges and enhance the effectiveness and accountability of trusted flaggers, a set of comprehensive recommendations have been proposed. Chief among them is the clear and explicit guidelines, along with delegated acts, can offer certainty and insight, fostering coordination and cooperation among all stakeholders. These guidelines can offer insight, providing consistency among different member states and online platforms for a harmonized application of the regulation. European Commission could create these guidelines and delegated acts to provide consistency among the member states and for a solid legal ground for enforcing the obligations of trusted flaggers. A multi-stakeholder forum consisted of representatives of different groups like online platforms, trusted flaggers, NGOs and national authorities could be created for the commission to consult and have feedback. This could foster a participatory and collaborative approach and provide a balanced and comprehensive solution which can address the legal and practical aspects of trusted flaggers.

Moreover, ensuring the success of the DSA's measures requires fostering effective cooperation and coordination among diverse actors, including online platforms, Trusted Flaggers, competent authorities, Digital Services Coordinators, and the European Board for

Digital Services. Addressing challenges related to communication, information-sharing, consistency, and trust among these actors is vital in fostering a safer online environment.

This thesis filled a gap in the literature by (*i*) performing a compare-contrast analysis of the previous regulatory regimes for content moderation with the DSA to reveal deficiencies in the online content moderation practices, (*ii*) applying the trusted flaggers system for the content moderation to explore distinctive consequences for freedom of expression and (*iii*) investigating how these frameworks can be enhanced to remain relevant while providing a safer and accountable digital landscape. Consequently, this study makes it easier to consider a balanced regulatory approach to mitigate freedom of expression-related issues and risks in a holistic manner.

On the other hand, the present thesis is subject to certain limitations, which in turn suggest areas for future research: (*i*) This study presents findings on the content moderation practices through an analysis of the previous self-initiated practices of the platforms and the rules set by the ECD, (*ii*) Research was made based on the system of trusted flaggers as introduced in the DSA while comparing it with similar notifiers; as a result, the topic in question has been interpreted in light of current practices, (*iii*) The DSA came into force, but it has not been implemented for a long time. Therefore, the results in practice are yet to be seen.

Considering all the factors and insights, the implementation of the trusted flaggers system for content moderation under the DSA is a multifaceted undertaking that necessitates a thorough approach. Striking a balance between addressing illegal content and safeguarding freedom of expression at the same time demands adherence and attention to human rights principles. By embracing the recommendations of creating ex-ante guidelines, providing effective cooperation and coordination among all actors involved in content moderation process, and involving trusted flaggers in the design and implementation of content moderation while addressing challenges proactively, the DSA can achieve its ultimate goal of fostering a safer and more accountable online environment while protecting users' rights and promoting responsible online speech. Trusted flaggers can be implemented effectively by ensuring and regularly checking that they meet the criteria set by the DSA. The implementation of trusted flaggers in this whole process will not only enhance content moderation but also reinforce a digital landscape that respects and upholds fundamental human rights in this evolving digital age. Through collective effort and coordination among the actors, the potential of trusted flaggers could be realized. The DSA is establishing provisions for this collective effort and coordination, but it also has some room for

improvement and clarification. The implementation of the rules set can be improved by specifying the details of how online platforms should prioritize the notices they receive from trusted flaggers, and providing clear rules on what constitutes illegal content rather than leaving it to national laws of the member states for consistent implementation of the rules in the EU. This way, a more secure and responsible online environment can be fostered by successfully integrating trusted flaggers within the DSA's framework, upholding both accountable content oversight and the preservation of fundamental human rights.

BIBLIOGRAPHY

PRIMARY SOURCES

LEGISLATION

Primary EU Law

European Union Charter of Fundamental Rights of the European Union [2012] OJ C 326

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

Secondary EU Law

Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU 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) in view of changing market realities, *OJ L 303, 28.11.2018*

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce') OJ L 178, 17.7.2000

Directive 2011/93/EU of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography and replacing Council Framework Decision 2004/68/JHA *OJ L 335, 17.12.2011*

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC

The Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC

Other European Institutions Documents

Code of Conduct on Countering Illegal Hate Speech Online

Commission Staff Working Document ‘Online services, including e-commerce, in the Single Market’, SEC (2011) 1641 (final) accompanying COM (2011) 942

Commission Staff Working Document, 'Executive Summary of the Impact Assessment Report Accompanying the document Proposal For A 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 {COM(2020) 825 final} - {SEC(2020) 432 final} - {SWD(2020) 348 final}'

Communication from The Commission to The European Parliament, The Council, The European Economic And Social Committee And The Committee Of The Regions Tackling Illegal Content Online towards an enhanced responsibility of online platforms (2017) COM/2017/0555 final

Council of Europe, 'Content Moderation: Best practices towards effective legal and procedural frameworks for self-regulatory and co-regulatory mechanisms of content moderation' May 2021, https://rm.coe.int/content-moderation-en/1680a2cc18

Council of Europe, 'Guidance Note on Content moderation' 20 September 2021 https://www.coe.int/en/web/freedom-expression/-/guidance-note-on-content-moderation

EU Code of Conduct on countering illegal hate speech online – European Commission, May 2016, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

European Commission, 'Communication from the Commission to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions: Tackling Illegal Content Online Towards and Enhanced Responsibility of Online Platforms', COM/2017/0555 final

First Report on the application of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market, COM (2003) 702, 21.11.2003

Legislation from other jurisdictions

German Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG) Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, 1 September 2017

CASE LAW

Case C-236/08, *Google France, Google Inc v Louis Vuitton Malletier SA and Others* [2010], ECLI:EU:C:2010:159

Delfi AS v. Estonia, Application no. 64569/09, (2015) EHRR

Magyar Tartalomszolgáltatók Egyesülete and Index.Hu Zrt v. Hungary, Application no. 22947/13, (2016) EHHR

SECONDARY SOURCES

BOOKS

Aina Turillazzia, Mariarosaria Taddeo, Luciano Floridia, Federico Casolari, 'The digital services act: an analysis of its ethical, legal, and social implications', LAW, INNOVATION AND TECHNOLOGY, 2023, VOL. 15, NO. 1, 83–106 (10 Oct 2022) https://doi.org/10.1080/17579961.2023.2184136

Aleksandra Kuczerawy, 'Intermediary liability & freedom of expression: Recent developments in the EU notice & action initiative' (2015) https://doi.org/10.1016/j.clsr.2014.11.004

Anastasia Kozyreva, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser & Jason Reifler, 'Resolving content moderation dilemmas between free speech and harmful misinformation' 9 November 2022, https://www.pnas.org/doi/10.1073/pnas.2210666120, accessed 3 July 2023

Berrak Genç-Gelgeç, 'Regulating Digital Platforms: Will the DSA Correct its Predecessor' s Deficiencies?' (2022) 18 CYELP 25 https://www.cyelp.com/index.php/cyelp/article/view/485 accessed 9 August 2023

Gehan Gunatilleke, 'Justifying Limitations on the Freedom of Expression' 1 November 2020, https://link.springer.com/article/10.1007/s12142-020-00608-8, accessed 4 July 2023

Mark D. Cole, Christina Etteldorf and Carsten Ullrich, 'Updating the Rules for Online Content Dissemination: Legislative Options of the European Union and the Digital Services Act Proposal' 19 May 2021, http://dx.doi.org/10.5771/9783748925934 accessed 7 June 2023

Martin Husovec, 'The DSA as a Creator's Charter?' *Journal of Intellectual Property Law & Practice, 2023, Vol. 18, No. 2*

Naomi Appelman & Paddy Leerssen, 'On "Trusted" Flaggers', Yale Journal of Law and Technology vol.24, 452

Sebastian Felix Schwemer, 'Trusted Notifiers and the Privatization of Online Enforcement' *Computer Law & Security Review, Volume 35, Issue 6,* (2019), https://doi.org/10.1016/j.clsr.2019.105339 accessed 28 November 2022

Thiago Dias Oliva, 'Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression' Human Rights Law Review 20, 5 December 2020, p.614, https://doi.org/10.1093/hrlr/ngaa032 accessed 29 June 2023

Toygar Hasan Oruç, 'The Prohibition of General Monitoring Obligation for Video-Sharing Platforms under Article 15 of the E-Commerce Directive in light of Recent Developments: Is it still necessary to maintain it?' JIPITEC 13 (3) 2022, https://www.jipitec.eu/issues/jipitec-13-3-2022/5555 accessed 11 May 2023

ARTICLES

Folkert Wilman, 'The Digital Services Act (DSA) - An Overview' (16 Dec 2022), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4304586 accessed 18 May 2023

Heidi Tworek & Paddy Leerssen, 'An Analysis of Germany's NetzDG Law' Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression, 2019, https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf accessed 2 July 2023

Christian Alberdingk Thijm, '10 Questions about the Digital Services Act, 2022, https://www.bureaubrandeis.com/10-questions-about-the-digital-services-act/ accessed 21 November 2022

Ilaria Buri and Joris van Hoboken, 'The Digital Services Act (DSA) proposal: a critical overview', Digital Services Act (DSA) Observatory, Institute for Information Law (IViR), University of Amsterdam, Discussion Paper (28 October 2021) https://dsa-observatory.eu/wp-content/uploads/2021/11/Buri-Van-Hoboken-DSA-discussion-paper-Version-28_10_21.pdf accessed 15 May 2023

Miriam C. Buiten, 'The Digital Services Act: From Intermediary Liability to Platform Regulation', 21 June 2021http://dx.doi.org/10.2139/ssrn.3876328 accessed 20 November 2022

Rachel Griffin, 'New School Speech Regulation and Online Hate Speech: A Case Study of Germany's NetzDG' SSRN Electronic Journal 2021 https://sciencespo.hal.science/hal-03586791/document accessed 1 July 2023

Raphaël Gellert and Pieter Wolters, 'The revision of the European framework for the liability and responsibilities of hosting service providers: Towards a better limitation of the dissemination of illegal content' (7 April 2021), page 28, https://www.eerstekamer.nl/eu/documenteu/_the_revision_of_the_european/f=/vlmulhixjsbs.pdf accessed 15 May 2023

Teresa Rodríguez de las Heras Ballell, 'The background of the Digital Services Act: looking towards a platform economy Teresa Rodríguez de las Heras Ballell' ERA Forum (2021) https://link.springer.com/article/10.1007/s12027-021-00654-w accessed 22 November 2022

OTHER SOURCES

Aleksandra Kuczerawy, 'The Good Samaritan that wasn't: voluntary monitoring under the (draft) Digital Services Act' Verfassungsblog on Matters Constitutional (2021) https://verfassungsblog.de/good-samaritan-dsa/ accessed 23 November 2022

Alexandre De Streel et al., Center on Regulation in Europe, 'Online Platforms' Moderation of Illegal Content Online' (June 2020) https://cerre.eu/news/study-online-platforms-moderation-of-illegal-content-online/ accessed 10 May 2023

Alexandre De Streel & Martin Husovec, 'The e-commerce Directive as the cornerstone of Internal Market: Assessment and Options for Reform' European Parliament Policy Department for Economic, Scientific and Quality of Life Policies May 2020, https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2020)648797 accessed at 11 August 2021

Anja Hoffmann & Alessandro Gasparotti, 'Liability for illegal content online: Weaknesses of the EU legal framework and possible plans of the EU Commission to address them in a "Digital Services Act"' (March 2020) https://www.cep.eu/fileadmin/user_upload/hayek-stiftung.de/cepStudy_Liability_for_illegal_content_online.pdf accessed

Brussels Report, 'The DSA is nothing more than digital censorship' (2022) https://www.brusselsreport.eu/2022/01/20/the-dsa-is-nothing-more-than-digital-censorship/ accessed 23 November 2022

Claude-Étienne Armingaud, Dr. Ulrike Elteste, Camille J. Scarparo, Dr. Thomas Nietsch, Andreas Müller 'EU Digital Services Act: Fundamental Changes for Online Intermediaries' (11 April 2022) https://www.klgates.com/eu-digital-services-act-fundamental-changes-for-online-intermediaries-11-4-2022 accessed 13 March 2023

Christopher Herwartz 'Now there's finally an answer to the destructive power of social media' https://www.handelsblatt.com/meinung/kommentare/kommentar-jetzt-gibt-es-endlich-eine-antwort-auf-die-zerstoererische-kraft-der-sozialen-medien/27981470.html?ticket=ST-1888944-JCAOkcZgsvjjp1MOpff4-ap1 accessed 24 November 2022

CMS Law Now, 'The E-commerce Directive' (12 April 2002) https://cms-lawnow.com/en/ealerts/2002/04/the-e-commerce-directive?format=pdf&v=4 accessed 10 May 2023

European Commission, 'The Digital Services Act: ensuring a safe and accountable online environment: What are the key goals of the Digital Services Act?' https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#what-are-the-key-goals-of-the-digital-services-act accessed 1 May 2023

European Commission, 'e-Commerce Directive' Shaping Europe's Digital Future, June 2022, https://digital-strategy.ec.europa.eu/en/policies/e-commerce-directive accessed 11 August 2023

European Commission, 'The Digital Services Act Package' (2023) https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed 12 August 2023

EDRi, 'The Digital Services Act, The EDRi guide to 2,297 amendment proposals' October 2021, https://edri.org/wp-content/uploads/2021/10/EDRi-policy-paper-Digital-Services-Act-Nov-2021.pdf accessed 11 June 2023

Eliska Pirkova, 'The Digital Services Act: your guide to the EU's new content moderation rules' 6 July 2022, https://www.accessnow.org/digital-services-act-eu-content-moderation-rules-guide/ accessed 9 July 2023

European Commission, 'Digital Services Act: Questions and Answers' (24 April 2023) https://digital-strategy.ec.europa.eu/en/faqs/digital-services-act-questions-and-answers accessed 10 May 2023

European Commission, 'e-Commerce Directive' https://digital-strategy.ec.europa.eu/en/policies/e-commerce-directive accessed 10 May 2023

European Commission, 'The Digital Services Act: ensuring a safe and accountable online environment: What are the key goals of the Digital Services Act?' https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en#what-are-the-key-goals-of-the-digital-services-act accessed 1 May 2023

European Commission, 'Digital Services Act: EU's landmark rules for online platforms enter into force' (2022) < https://ec.europa.eu/commission/presscorner/detail/en/IP_22_6906> accessed on 21 November 2022

European Commission, 'Questions and Answers: Digital Services Act' (2022) https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2348 accessed 20 November 2022

European Data Protection Supervisor, 'EDPS orders Europol to erase data concerning individuals with no established link to a criminal activity' 10 Jan 2022, https://edps.europa.eu/press-publications/press-news/press-releases/2022/edps-orders-europol-erase-data-concerning_en#:~:text=On%203%20January%202022%2C%20the,EDPS'%20inquiry%20launched%20in%202019. accessed 09 June 2023

European Telecommunications Network Operators' Association, 'A clean and open Internet: Public consultation on procedures for notifying and acting on illegal content hosted by online intermediaries' https://www.etno.eu/datas/positions-papers/2012/etnoc01-dsm-notice-and-action-consultation-sep-2012.pdf accessed 15 May 2023

IMCO Committee, 'Online Platforms' Moderation of Illegal Content', (2020), https://www.europarl.europa.eu/RegData/etudes/STUD/2020/652718/IPOL_STU(2020)652718_EN.pdf accessed 21 November 2022

Iverna McGowan, 'How can we apply human rights due diligence to content moderation? Focus on the EU Digital Services Act – Event Summary' September 9, 2021 https://cdt.org/insights/how-can-we-apply-human-rights-due-diligence-to-content-moderation-focus-on-the-eu-digital-services-act-event-summary/ accessed 7 July 2023

Janosch Delcker, 'Germany's balancing act: Fighting online hate while protecting free speech' 1 October 2020, https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/ accessed 2 July 2023

Jen Carter, Growing Our Trusted Flagger Program Into YouTube Heroes, 22 September 2016, https://blog.youtube/news-and-events/growing-our-trusted-flagger-program/ accessed 29 June 2023

Joan Donovan 'Why Social Media Can't Keep Moderating Content in the Shadows' (2020) MIT Technology Review https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/ accessed 24 November 2022

Joseph Downing 'The EU's Digital Services Act: Europeanising social media regulation?' (2022) LSE Comment https://blogs.lse.ac.uk/europpblog/2022/08/08/the-eus-digital-services-act-europeanising-social-media-regulation/ Accessed 23 November 2022

Kerstin Bäcker, 'The trusted flaggers in the Digital Services Act', 23 May 2022, https://www.lausen.com/en/the-trusted-flaggers-in-the-digital-services-act/ accessed 11 June 2023

Konstantinos Komaitis, Katitza Rodriguez and Christoph Schmon, 'Enforcement Overreach Could Turn Out To Be A Real Problem in the EU's Digital Services Act' Electronic Frontier Foundation, 18 February 2022 https://www.eff.org/deeplinks/2022/02/enforcement-overreach-could-turn-out-be-real-problem-eus-digital-services-act accessed 10 June 2023

Meta, 'How Meta's third-party fact-checking program works' 1 June 2021 https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works accessed 29 June 2023

Natasha Duarte, Emma Llanso and Anna Loup, 'Mixed Messages? The Limits of Automated Social Media Content Analysis' *Center for Democracy & Technology,* Proceedings of Machine Learning Research 81:1-1 2018 https://proceedings.mlr.press/v81/duarte18a.html accessed 4 July 2023

Nicolo Zingales, Thales Bertaglia and Catalina Goanta, 'Freedom of expression & Content Moderation on Social Media: The Promises and Challenges of Experimental Accountability' Maastricht University Blog, 10 August 2021, https://www.maastrichtuniversity.nl/blog/2021/08/freedom-expression-content-moderation-social-media-promises-and-challenges-experimental accessed 3 July 2023

Sally O'Brien, 'E-Commerce Directive versus the new Digital Services Act: is there a new liability regime for online service providers?', 2 November 2022, https://www.loganpartners.com/e-commerce-directive-versus-the-new-digital-services-act-is-there-a-new-liability-regime-for-online-service-providers/ accessed on 25 November 2022

Sara Ataei, Jens Van Lathem and Roeland Moeyersons, 'The Practical Implications of the Regulations on Digital Services and Markets' 08 May 2023, https://seeds.law/en/news-insights/the-practical-implications-of-the-regulations-on-digital-services-and-markets/ accessed 10 June 2023

Sebastian Becker Castellaro & Jan Penfrat, 'The DSA fails to reign in the most harmful digital platform businesses – but it is still useful' , Verfassungsblog on Matters Constitutional, (2022) https://verfassungsblog.de/dsa-fails/ accessed 24 November 2022

Stefan Theil, 'The German NetzDG: A Risk Worth Taking?' 2018 https://verfassungsblog.de/the-german-netzdg-a-risk-worth-taking/ accessed 2 July 2023

Tambiama Madiega, European Parliamentary Research Service, 'Digital Services Act' (17 November 2022) https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)689357 accessed 11 May 2023

Tremau, 'New Role of Trusted Flaggers In The EU', 25 May 2022, https://tremau.com/digital-services-act-trusted-flagger-organisations accessed 10 June 2023

United Nations Human Rights Office of the High Commisioner, 'Moderating online content: fighting harm or silencing dissent?', 23 July 2021, https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent> accessed 12 March 2023

Valentina Golunova, The Digital Services Act and freedom of expression: triumph or failure? (2021) Maastricht University Blog <https://www.maastrichtuniversity.nl/blog/2021/03/digital-services-act-and-freedom-expression-triumph-or-failure> accessed 22 November 2022

Will Oremus, 'How social media 'censorship' became a front line in the culture war' (9 October 2022) <https://www.washingtonpost.com/technology/2022/10/09/social-media-content-moderation/> accessed 12 March 2023

YouTube, 'About the YouTube Priority Flagger program' https://support.google.com/youtube/answer/7554338?hl=en#zippy= accessed 11 June 2023