

THE INTERPLAY BETWEEN ONLINE HATE SPEECH AND INTERMEDIARY LIABILITY IN EU AND UK LAW

Department: School of Law and Social Science, Oxford Brookes University

Name: Fabiana Ciccarella

Student ID: 19138299

Words count: 46,326

Abstract

This research examines strategies adopted by Internet intermediaries to counteract the proliferation and impact of hate speech on social media platforms while safeguarding the freedom of expression. It scrutinises the current landscape of hate speech within the UK and Europe, analysing existing legislative frameworks and their approaches to objectionable content. Emphasis is placed on evaluating the compatibility of these legal provisions with essential human rights, notably the right to freedom of expression. The study highlights the Digital Services Act (DSA) as a recent European legislative initiative aimed at establishing a robust regulatory framework for online content, including provisions for user-initiated reporting and platform-based content review processes. Additionally, it delineates a risk-based approach to content moderation, aligning with overarching objectives of user protection and societal welfare. Conversely, the UK's 2023 Online Safety Act (OSA) emphasises holding online service providers accountable for illegal and harmful content, although gaps in content moderation measures and transparency are identified. Recommendations to policy makers include broadening the scope of harmful content definition to encompass cyberbullying and online harassment. Ultimately, the study asserts the need for both legislative measures to carefully reconcile the imperatives of content moderation with the preservation of freedom of expression in the online domain.

Table of Contents

Chapter 1: Introduction	5
1.1. Outline of the problem	5
1.2. Legislative framework and terminology in a nutshell.....	7
1.3. Structure of the dissertation.....	12
Chapter 2: Hate Speech and Freedom of Speech Online: A Legal Analysis From The Perspective of the Liability Of Internet Service Providers	14
2.1 Introduction	14
2.2 Exceptionalism in regulating online hate speech	20
2.3 Balancing Platform Liability and Fundamental Rights in the digital age: The E-commerce Directive.....	23
2.4 Is there a definition of on online hate speech? The CoE legislative framework.....	29
2.5 Evolution and Critique of European Union Regulatory Approaches to Online Hate Speech: From Voluntary Codes (Soft Law) to Hard-Law Measures	31
2.6 The Jurisprudential Dynamics of Hate Speech Liability in Online Platforms within the European Court of Human Rights (ECHR) in Strasbourg and the Court of Justice of the European Union (CJEU)	37
2.7 The Evolving Role of the CJEU in Shaping Intermediary Liability and Freedom of Expression Online.....	48
2.8 The European Court of Human Rights and the Council of Europe on the responsibility of intermediaries and the protection of human rights in the digital world.....	54
2.9 Conclusion.....	56
Chapter 3: Assessing the “Systemic Risks” in Digital Services Act and International Human Rights Law	58
3.1 Introduction	59
3.2 The provisions of Digital Services Act in response to hate speech.....	63
3.3 The First Systemic Risk: Dissemination of Illegal Content – Defining and Addressing 'Illegal Content' in the DSA	73
3.4 The Second Systemic Risk: The Actual or Foreseeable Negative Effects on the Exercise of Fundamental Rights	78
3.5 Third and Fourth Categories of Systemic Risks: Manipulation of Very Large Online Platforms (VLOPs) and Search Engines - Effects and Consequences.....	81
3.6 Mitigation of systemic risks	83
3.7. International human rights law (IHRL) as a framework for the moderation of online content	86
3.8. Hate speech and International Human Rights Law	87
3.9 Conclusions	90
Chapter 4: Regulatory Approaches to Online Harm. A Comparative Analysis of European and English Legislation	92
4.1 Introduction	92

4.2	The scope and limitations of the Online Safety Act: Balancing Regulation and Innovation in Platform Governance'.....	96
4.3	Revisiting Online Harm: Doctrinal Debates on Online Safety Regulatory Approaches ..	102
4.4	Navigating the complexity of harmful content regulation in the Online Safety Act	108
4.5	Platform accountability: a comparison between the 'duty of care' in the OSA and the 'good Samaritan' rule in the DSA	112
4.6	Conclusions	120
Chapter 5: Conclusion and Recommendations		123
5.1	Conclusion.....	123
5.2	Recommendations	125
Bibliography		128

Chapter 1: Introduction

1.1. Outline of the problem

Social media platforms have become breeding grounds for hate speech and harmful content. The rapid advancement of technology characteristic of the current era has accelerated the dissemination of expressions of hatred, perpetuation of atrocities against marginalised groups, instigation, and bias¹. The accumulation of disparaging online commentary has the potential to foment violence, exclude minority groups, and gradually undermine the bedrock of an inclusive societal framework². As a result, this limited communication environment poses a threat to social cohesion and can distort public perception of ongoing discussions. At its core, this dilemma presents a nuanced balance between upholding the fundamental principle of free expression, crucial for the vitality of democracy, and mitigating the adverse effects of hate speech on individuals³. In pursuit of bolstering international oversight in the regulation of hate speech and content moderation, the European Union (EU) concluded an accord with prominent tech entities, including YouTube, Twitter, Microsoft, and Facebook, in 2016. This agreement entailed the formulation of a code of conduct aimed at curtailing the proliferation of hate speech across online platforms⁴. The platforms committed to following community guidelines and regulations outlined in this code, making necessary adjustments to their align with these standards. However, in practice, adherence has been inconsistent. Moreover, the implementation of the Digital Services Act (DSA) in the EU in 2022 strengthens the existing content moderation framework by enhancing the internal complaint handling mechanisms of these platforms

¹ Akbarova, Lala. "Hate Speech-Legal, Political, Cultural Aspects. World Experience In Legal Regulation." *Distance Education in Ukraine: Innovative, Normative-Legal, Pedagogical Aspects* 2, pp.22-44

² Gorenc, Nina. "Hate speech or free speech: an ethical dilemma?" *International Review of Sociology*, 2022, pp.413-425

³ De Moraes, Carlos Blanco. "'Digital Democracy': A Threat to the Democratic System or Oxygenation of Representative Democracy and Free Speech?" *The Rule of Law in Cyberspace*. Cham: Springer International Publishing, 2022, pp. 9-63.

⁴European Commission. "The EU Code of Conduct on Countering Illegal Hate Speech Online." 2016.

of these platforms. In response to concerns regarding the limitations of prevailing content moderation systems employed by social media platforms, the DSA presents a comprehensive regulatory framework aimed at addressing hate speech, albeit with certain limitations⁵. In the UK, the Online Safety Act (OSA) was introduced to ensure online platforms take proactive measures to protect users from harmful and illegal content, including hate speech, by imposing a statutory ‘duty of care’ and requiring transparent reporting and accountability mechanisms. Against this backdrop, this work focuses on the complex interplay between freedom of expression, hate speech, and the challenges faced by both governments and social media platforms in addressing the issues. This study aims to provide a comprehensive analysis of the Digital Services Act (DSA) and the UK Online Safety Act (OSA), with a particular emphasis on their impact on digital service providers (DSPs) and the broader implications for fundamental rights. By providing detailed recommendations, this study advocates for the creation of digital content regulations that are both balanced and effective, while ensuring respect for fundamental rights. The digital realm is faced with the tension between safeguarding free speech and mitigating the harms of hate speech. Online platforms, also known as intermediaries, struggle to achieve this delicate balance. The imperative to achieve equilibrium is further intensified by growing user frustrations when their content is removed from platforms due to allegations of hate speech⁶. Additionally, users often lack sufficient understanding of what constitutes hate speech, especially because there is no universally acceptable definition of hate speech⁷. Consequently, intermediaries encounter challenges in defining the extent of accountability they assume for the content they facilitate on their platforms. This notion of liability, termed intermediary liability, is further complicated by the diverse roles that intermediaries perform, ranging from ‘mere

⁵ Ortolani, Pietro. "The Digital Services Act, Content Moderation and Dispute Resolution." *Content Moderation and Dispute Resolution* 2023, pp.1-20.

⁶ Wilson, Richard Ashby, and Molly K. Land. "Hate speech on social media: Content moderation in context." *Conn. L. Rev.* 2020, pp.1034-1050.

⁷ MacAvaney, Sean, et al. "Hate speech detection: Challenges and solutions." *PloS one* 2019.

conduits' for information to 'content cache service providers' or comprehensive service providers⁸. In the following chapters, this research will examine and analyse case laws that illustrate various instances and interpretations of legal provisions concerning intermediaries and their responsibilities. Additionally, it will explore the efforts of courts to navigate the complex task of balancing the protection of freedom of expression with the mitigation of hate speech, a challenge that has proven to be particularly difficult.

1.2. Legislative framework and terminology in a nutshell

Historically, the regulation of intermediaries and online platforms has faced challenges in balancing the conflicting rights impacted by the activities of digital service providers (DSPs)⁹. Governments and rights holders have sought to engage these intermediaries in the removal of content from the internet that is deemed illegal¹⁰. However, transferring this duty from the public to the private sector has proven ineffective in reducing online hostility due to various limitations¹¹. Delays in removing content, particularly hate speech, can significantly affect internet users by exacerbating emotional distress, fostering an environment of hostility, and potentially inciting real-world violence and discrimination. In response, international bodies like the European Union (EU) have adopted a multi-pronged approach to combat hate speech on digital platforms¹². Over time, the legal landscape within the EU regarding online hate speech has evolved significantly due to the growing importance of the Internet and the challenges it presents, including hate speech and other criminal activities. Previously, online platforms were not held accountable for the content on their platforms unless they were aware of its illegal nature and failed to remove it promptly. However, inconsistencies among EU member states in dealing with online hate speech have highlighted the need for more uniform regulations. To

⁸Șomlea, Radu. "A Comparative Analysis of the EU and US Trends in Regulating Intermediary Service Provider Liability." *LL.M Short Thesis*, Central European University, 2017, pp 1-39.

⁹ Frosio, Giancarlo, and Christophe Geiger. "Taking fundamental rights seriously in the Digital Services Act's platform liability regime." *European Law Journal* 2023 pp. 31-45.

¹⁰ *Ibid.*, p.32.

¹¹ *Ibid.*, pp. 31-39.

¹² Leiser, Mark. "Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation." 2023, p.4.

address these issues, the EU introduced the E-Commerce Directive in 2000, aiming to harmonise national legal systems across the EU concerning intermediary liability. Under this Directive, online platforms were considered intermediaries and were not held accountable for the content published on their platforms unless they were aware of its illegal nature and neglected to remove it in a timely manner¹³. Initially, the regulatory framework for online content was limited, with each EU member state adopting distinct approaches to combat hate speech. Consequently, different Member States (MSs) have shown notable disparities across jurisdictions. With the spread of online hate speech, there has been an increasing demand for more consistent regulatory frameworks. For instance, the E-commerce Directive (2000/31/EC) aimed to address intermediary liability issues and required implementing measures to harmonise disparate national legal systems across the EU¹⁴. Recently, in 2024, the EU amended the E-Commerce Directive and introduced the Digital Service Act (DSA). This legislative framework seeks to implement a more robust and comprehensive approach to content governance, balancing enhanced user freedoms with the preservation of fundamental rights. Under the DSA, online platforms or intermediaries are required to conduct mandatory risk assessments before removing content suspected to be harmful. Additionally, the DSA mandates online platforms to establish transparent procedures for users to report harmful content and challenge content removal¹⁵. Another significant framework is provided by the Council of Europe (CoE). However, it is important to note that this thesis will not offer an in-depth analysis of the CoE legislation, as the primary focus of this research is on the EU and UK legal frameworks. The CoE has been instrumental in advancing human rights, primarily through its framework in the European Convention

¹³Bertolini, Andrea, Francesca Episcopo, and NICOLETA-ANGELA Cherciu. "Liability of online platforms", 2021, pp. 28-48.

¹⁴ Lodder, Arno R. "Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the internal market." *EU Regulation of E-Commerce*. Edward Elgar Publishing, 2017, pp. 15-30.

¹⁵ Buri, Ilaria, and Joris van Hoboken. "The Digital Services Act (DSA) proposal: a critical overview." *Digital Services Act (DSA) Observatory*, 2021, pp.25-31.

for Human Rights (ECHR) and the jurisprudence of the European Court of Human Rights (ECtHR)¹⁶. This role has been influential in shaping national strategies to tackle online hate speech. Furthermore, the CoE holds member states accountable by monitoring their compliance with standards and publishing reports on hate speech. Various judicial precedents embracing this notion will be analysed to delineate the boundaries and balance between permissible expressions and hate speech. In their analysis of the case law concerning freedom of expression on the internet, Mir, Joan Barata, and Marco Bassini highlight the significant roles played by both the Court of Justice of the European Union (CJEU) and the European Court of Human Rights (ECtHR) in balancing the right to freedom of expression with the need to combat hate speech¹⁷. The Court of Justice of the European Union (CJEU) and the ECtHR are pivotal institutions instrumental in harmonising the right to freedom of expression with efforts to combat hate speech. Their decisions significantly shape contemporary legal frameworks. The pronouncements of the CJEU can help broaden the scope of the regulatory framework governing online platforms and clarify its limitations. Conversely, the ECtHR examines whether national legislation, including laws implementing EU directives, is compatible with the ECHR¹⁸. Therefore, the jurisprudence of this Court on freedom of expression and hate speech, and its limitations, serves as a reference point for both national courts and governmental bodies¹⁹. The ECtHR has developed a nuanced assessment framework to balance the freedom of expression with other interests²⁰. The first tier requires assessing whether there is a 'legitimate aim', meaning any

¹⁶Cliquennois, Gaëtan, Sonja Snacken, and Dirk van Zyl Smit. "Can European human rights instruments limit the power of the national state to punish? A tale of two Europes." *European Journal of Criminology*, 2021 pp. 14-22.

¹⁷ Mir, Joan Barata, and Marco Bassini. "Freedom of expression in the internet: main trends of the case law of the European Court of Human Rights." *The Internet and Constitutional Law*. Routledge, 2016, pp. 75-88.

¹⁸Grabenwarter, Christoph. "The European Convention on Human Rights: Inherent Constitutional Tendencies and the Role of the European Court of Human Rights." *ELTE LJ* (2014), pp.101-103.

¹⁹ Mchangama, Jacob, and Natalie Alkiviadou. "Hate speech and the European Court of Human Rights: Whatever happened to the right to offend, shock or disturb?", *Human Rights Law Review* 2021, p. 13.

²⁰ Voorhoof, D., "The Right to Freedom of Expression and Information under the European Human Rights System: Towards a More Transparent Democratic Society," *RSCAS 2014/12* (European University Institute, 2014), pp. 1-13.

restriction on freedom of expression must be based on a justifiable rationale, such as protecting the rights of individuals or ensuring public safety²¹. The second tier involves assessing the ‘necessity in a democratic society’, which mandates that restrictions must be proportionate to the legitimate aim pursued and be the least intrusive means to achieve that aim. Thus, restrictions should not exceed what is essential to protect individual rights or uphold public safety²². In the United Kingdom (UK), the enactment of the Online Safety Act 2023 was introduced to tackle detrimental content, including hate speech²³. The Act’s oversight is entrusted to Ofcom, the UK's communications regulator. Ofcom, an independent body with statutory powers, is responsible for regulating broadcasting, telecommunications, and postal industries. Its role in overseeing the Online Safety Act includes ensuring that online platforms comply with the statutory duty of care to protect users from harmful and illegal content. Ofcom has discretionary powers to enforce the Act, which includes imposing fines and other penalties on platforms that fail to adhere to its regulations²⁴. The UK's approach to regulating online content also relies heavily on its common law tradition, where judicial precedent plays a crucial role in defining the parameters of permissible and impermissible speech. This tradition ensures that the interpretation and application of laws evolve through case law, which complements statutory regulations like the OSA by providing a nuanced and adaptable legal framework. Additionally, it is important to highlight that this thesis advocates for the adoption of International Human Rights Law (IHRL) as the most effective framework for moderating online content. While this thesis posits that IHRL offers the best balance between protecting freedom of expression and

²¹ Voorhoof, D., "The Right to Freedom of Expression and Information under the European Human Rights System: Towards a More Transparent Democratic Society," *RSCAS 2014/12* (European University Institute, 2014), pp. 1-13.

²² *Ibid.*, pp. 1-13

²³ Coe, Peter. "The Draft Online Safety Bill and the regulation of hate speech: have we opened Pandora’s box?" *Journal of Media Law* 14.1 (2022), pp. 50-75.

²⁴ Ofcom, the Office of Communications, is the UK’s independent regulatory authority overseeing broadcasting, telecommunications, and postal services. It ensures compliance with standards, promotes competition, protects consumers, and manages the radio spectrum. Additionally, under the Online Safety Act, Ofcom regulates online content to safeguard users, particularly children and vulnerable individuals, from harmful and illegal material. Established by the Communications Act 2003, Ofcom operates independently from the government to ensure impartiality.

regulating harmful content, it does not provide an exhaustive analysis of this perspective. This is one of the limitations of the thesis, alongside the limited in-depth analysis of CoE legislation. In the following chapters, this work will identify different instances and interpretations of legal provisions concerning intermediaries and their responsibilities. It will also explore Courts' efforts to reconcile freedom of expression with the mitigation of hate speech, a task that has proven challenging.

Key terms such as hate speech, harm, risk, and freedom of expression are hereby defined to facilitate a clear understanding of the discussions that follow:

- **Hate Speech** refers to any kind of communication in speech, writing, or behaviour that attacks or uses pejorative or discriminatory language toward an individual or a group based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factors²⁵.

- **Harm** in the context of online content regulation refers to the adverse effects that hate speech can have on individuals and society, including psychological distress, incitement to violence, and broader societal impacts such as the erosion of social cohesion and the perpetuation of discrimination and prejudice²⁶.

- **Risk** within content moderation, risk pertains to the potential for online content to cause harm. This involves assessing the likelihood and severity of negative outcomes resulting from the dissemination of harmful content²⁷.

- **Freedom of Expression** a fundamental human right allowing individuals to express themselves without interference. Enshrined in various international human rights documents, including Article 10 of the European Convention on Human Rights (ECHR), this right is not absolute and can be restricted when balancing other rights and societal interests, such as preventing hate speech²⁸.

²⁵Council of Europe, Committee of Ministers Recommendation No. R (97) 20

²⁶European Commission. (2016). "Code of Conduct on Countering Illegal Hate Speech Online".

²⁷Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' Plataforma por la Libertad de Infoacion, pp.3-24.

²⁸Council of Europe. (1950). *European Convention on Human Rights*. Article 10.

By grounding our discussion in these definitions, we can navigate the complex legal terrain of online hate speech and intermediary liability more precisely.

1.3. Structure of the dissertation

The dissertation is organised into five chapters, each addressing different facets of online hate speech and intermediary liability within the frameworks of the EU and the UK law. The first chapter explores the phenomenon of online hate propagated through social media platforms managed by Digital Service Providers (DSPs). The second chapter provides a detailed legal analysis of online hate speech and online freedom of speech, with a particular focus on the European Court of Human Rights (ECtHR) and its jurisprudence. This chapter examines how the ECtHR balances the protection of freedom of expression with the need to mitigate hate speech. The third chapter, divided into two parts, assesses the ‘systemic risks’ in the DSA and proposes International Human Rights Law (IHRL) as the standard to adopt for moderating online content. The first part focuses on how the DSA respond to hate speech, defining and addressing ‘illegal content’, the impact on fundamental rights, the manipulation of Very Large Online Platforms (VLOPs) and search engines, and measures to mitigate these risks. The second part emphasises the adoption of IHRL as the framework for moderating online content, highlighting its importance in balancing the protection of freedom of expression with the regulation of harmful content. The fourth chapter investigates the provisions of the UK Online Safety Act (OSA) concerning hate speech and harmful content, focusing on its regulatory impact and comparing it with the DSA. Finally, the fifth chapter offers final remarks and recommendations to policymakers, summarising the findings and addressing identified research gaps. To comprehensively explore the legal frameworks governing online hate speech in the EU and the UK, this dissertation employs documentary analysis and legal doctrinal scrutiny²⁹. Documentary analysis entails meticulous examination of existing documents and pertinent published materials relevant to the research inquiry: “How do the legal frameworks of the EU and UK address the liability of internet

²⁹O’Connor, Margaret. "Documentary analysis and policy." *Research methods in palliative care* (2007), pp. 230-237.

intermediaries in regulating online hate speech while balancing the right to freedom of expression?”. This analysis includes examining relevant documents and published materials, such as treaties, EU directives (e.g., the E-Commerce Directive and DSA), domestic statutes like the UK OSA, and case law from the CJEU and ECtHR. Additionally, the dissertation analyses academic literature, including scholarly journals, books and related works on freedom of expression, hate speech, intermediary liability, and human rights law. The legal doctrinal analysis focuses on interpreting and applying legal principles and doctrines articulated in legislation and case law. This methodological approach involves identifying pivotal legal concepts, evaluating jurisprudential precedents, and assessing legal arguments to provide a thorough understanding of how these legal frameworks regulate online hate speech while balancing the right to freedom of expression³⁰. Special attention is given to the jurisprudence of the European Court of Human Rights (ECtHR) and its role in balancing freedom of expression with the need to mitigate hate speech.

³⁰ Wendel, W. Bradley. "Explanation in legal scholarship: The inferential structure of doctrinal legal analysis." *Cornell L. Rev.* 96 (2010), p.1035-1038.

Chapter 2: Hate Speech and Freedom of Speech Online: A Legal Analysis From The Perspective of the Liability Of Internet Service Providers

Summary: 2.1. Introduction – 2.2 Exceptionalism in regulating online hate speech 2.3 Balancing Platform Liability and Fundamental Rights in the digital age: The E-commerce Directive– 2.4 Is there a definition of on online hate speech? The Council of Europe legislative framework - 2.5 Evolution and Critique of European Union Regulatory Approaches to Online Hate Speech: From Voluntary Codes (Soft Law) to Hard-Law Measures – 2.6 The Jurisprudential Dynamics of Hate Speech Liability in Online Platforms within the European Court of Human Rights (ECHR) in Strasbourg and the Court of Justice of the European Union (CJEU)– - 2.7 The Evolving Role of the Court of Justice of the European Union in Shaping Intermediary Liability and Freedom of Expression Online – 2.8 The European Court of Human Rights and the Council of Europe on the responsibility of intermediaries and the protection of human rights in the digital world - 2.9 Conclusions

2.1 Introduction

This Chapter deals with the digitisation of information access and expression facilitated by the widespread adoption of the internet, which concurrently presents challenges in combating instances of hate speech among its online users. Addressing the potential threats to the safety and dignity of users, necessitates a multifaceted approach to harmonising the issue of online hate speech within the overarching principle of freedom of expression. The objective of this investigation is to address the primary research question of this thesis: “How do the European Union (EU) and United Kingdom (UK) legal frameworks manage the responsibility of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression”? There is a tension in balancing online hate speech and the freedom of expression which arises from the need to protect individuals and groups from harmful, abusive, and discriminatory speech, while simultaneously upholding the fundamental right to free speech. For instance, the E-Commerce Directive has introduced a

fundamental principle termed ‘Safe Harbour’ to shield Internet Service Providers (ISPs)³¹. According to Article 14 of this Directive, ISPs are shielded from liability for content published by third parties, unless they possess knowledge of illegal content (such as hate speech) and fail to act upon it expeditiously. Nevertheless, complex issues regarding the governance of illegal information on the Internet and the role of private Internet intermediaries, particularly large online platforms, remain to be addressed³². The original goals of the E-Commerce Directive (ECD) — facilitating cross-border electronic commerce, establishing a legal framework for online services, clarifying liability for internet intermediaries, and harmonizing regulations across Member States — remain relevant. However, the European Commission has identified the need for further interventions due to advancements in digital services, legal fragmentation, and enforcement challenges. Measures such as clearer due diligence obligations and enhanced notice-and-action procedures aim to address the substantial growth and diversification in the use of digital services. This underscores the necessity for updates to effectively govern current internet dynamics. These initiatives are intended to tackle new legal inconsistencies and issues related to the enforcement and supervision of these services. To tackle these evolving challenges, the Digital Services Act (DSA), adopted in 2022, marks a transformation in the EU's regulatory approach. It aims to address incidents of illegal content, including hate speech, and assigns significant moderation responsibilities to platforms. Specifically, Article 34 of the DSA

³¹ “The E-commerce Directive introduces a safe harbour principle, under which three types of online intermediaries who host or transmit content provided by a third party are exempt from liability under certain conditions. 1. Mere conduit service providers (Art. 12) are exempt from liability when the service provider is only passively involved in the transmission of data (i.e. interconnection to the internet provided by traditional internet service providers and network operators), and, 2. Caching providers (Art. 13) are exempt from liability when they temporarily and automatically store data in order to make transmission more efficient (e.g. proxy server) and if several technical conditions for storing the information are met (e.g. local copy identical to original), and, 3. Hosting providers (Art. 14) are exempt from liability when those companies storing data for their users (e.g. webhosting) do not know that they host illegal activity or information and act expeditiously to remove or disable access to the illegal information.” Available at: [https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA\(2020\)649404_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2020/649404/EPRS_IDA(2020)649404_EN.pdf)

³² Schwemer, Sebastian Felix. "Digital Services Act: a reform of the E-Commerce Directive and much more." *Research handbook on EU internet law*. Edward Elgar Publishing, 2023, pp.1-3

mandates that very large online platforms (VLOPs) take measures to mitigate systemic risks, which include the dissemination of illegal content, the protection of public health, minors, civic discourse, as well as the detection and removal of harmful content³³. This involves implementing a risk-based approach to content moderation and other systemic measures to address these risks. This shift of responsibility from the state to private providers aims to curb the spread of harmful content but raises significant concerns about potential overreach and the suppression of legitimate expression. Content moderation, particularly when mandated by law, can lead to excessive censorship as platforms may err on the side of caution to avoid penalties, thereby stifling free speech. Moreover, reliance on automated systems for content moderation, which are prone to errors and biases, further exacerbates the risk of unjustly silencing voices. This approach can also create inconsistencies in enforcement and transparency issues, as the criteria for determining harmful content might vary across different platforms and jurisdictions. Hence, while the intention behind Article 34 of the DSA is to protect users, it necessitates careful consideration and safeguards to ensure that the right to freedom of expression is not unduly compromised. Similarly, the Council of Europe has taken a broad approach to preventing hate speech on social media platforms. For example, the Additional Protocol to the Council of Europe's (CoE) Budapest Convention on Cybercrime (2001) criminalises xenophobic and racist content, promoting individual responsibility among internet users to combat and diminish hate speech³⁴. In Europe, there has been a recent shift towards holding online platforms more accountable and implementing stricter regulations. Conversely, countries like the United States have a strong tradition of First Amendment protections, often vigorously defending the interests of tech giants and digital platforms in the context of free speech³⁵. While this approach aligns closely with national

³³ See Regulation 2022/2065/EU 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), 2022 OJ L 277/1.

³⁴ See Council of Europe. (2003). Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189). Strasbourg.

³⁵ Citron, D. K., & Wittes, B. (2017). "The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity." *Fordham Law Review*, 86(2), 401-424.

priorities, it often overlooks the potential harm to other fundamental rights posed by unregulated illicit online content. This aligns with the argument that Europe is effectively making private companies become enforcers of human rights and protecting the rights of individuals. Stressing this point underscores the importance of regulatory frameworks that empower private companies to take an active role in safeguarding fundamental rights online, ensuring a balanced approach to online content regulation. The EU has pursued a distinct trajectory, exemplified by the Digital Services Act (DSA), which represents a compromise by the European Union to balance various competing interests in the digital space. It preserves the existing liability exemption and safe harbour regime while introducing more structured and rigorous oversight of digital platforms' content moderation and operations³⁶. This move ostensibly aims at enhancing user protection, promoting transparency, and advancing broader societal objectives, as elaborated upon in the forthcoming analysis in Chapter Three. Various approaches have been applied to address hate speech online. For instance, content moderation tools such as human reviewers and algorithms are currently employed by major social media platforms³⁷. However, the use of algorithms has faced criticism due to controversies surrounding censorship and potential infringement of human rights compared to outright removal by human reviewers³⁸. Other proposals to address online hate speech are raised by scholars³⁹, who support the enactment of international legal frameworks to reduce and combat hate speech. As it will be discussed in Chapter Three of this thesis, we strongly advocate for the adoption of international human rights law (IHRL) as the foundational framework for content moderation on social media platforms. Even though, these

³⁶ See Regulation 2022/2065/EU 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), 2022 OJ L 277/1.

³⁷ Gongane, Vaishali U., Mousami V. Munot, and Alwin D. Anuse. "Detection and moderation of detrimental content on social media platforms: current status and future directions", *Social Network Analysis and Mining* 12.1 (2022), pp.1-41.

³⁸ Llansó, E. et al. Artificial Intelligence, Content Moderation, and Freedom of Expression, TWG: Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Netherlands. 2020, pp.6-8 <https://policycommons.net/artifacts/8986433/untitled/9875488/> (accessed on 28 May 2024).

³⁹ Mchangama, J., Alkiviadou, N., & Mendiratta, R. 'THOUGHTS ON THE DSA' 2011, pp.1-16.

platforms, being private entities, are not legally obligated to adhere to international human rights treaties, there is a growing consensus among distinguished scholars supporting this perspective⁴⁰. Evelyn Aswad posits that international law provides the most effective framework for protecting freedom of expression⁴¹. In a similar vein, Susan Benesch asserts that while IHRL cannot be directly implemented, it can still serve as a guiding framework for content moderation⁴². The European Court of Human Rights (ECtHR) assumes a pivotal role in elucidating the contours of freedom of speech and its constraints, particularly in the context of hate speech, within the framework of the European Convention on Human Rights (ECHR)⁴³. Through its jurisprudence, the ECtHR delineates nuanced criteria for assessing the permissibility of restrictions on freedom of expression, notably in cases involving manifestations of hate speech⁴⁴. Its adjudicatory function serves to harmonise standards across European jurisdictions, fostering coherence in the application of human rights principles while addressing the complex interplay between safeguarding individual liberties and combating discriminatory or inflammatory discourse. In essence, the ECtHR's adjudicative mandate contributes significantly to the evolution of a robust legal framework for democratic governance and the protection of fundamental rights within the European legal landscape, particularly concerning the regulation of hate speech. One illustrative instance highlighting the ECtHR's role in adjudicating the tension between hate speech and freedom of expression is evident in the case of *Lehideux and Isorni v. France*. Here, the ECtHR ruled in favour of the applicants, François Lehideux and Jacques Isorni, finding that while their expression was indeed offensive, it did not constitute incitement to violence. The Court emphasised those certain historical facts, such as the Holocaust, are beyond the realm of protected expression. This suggests that, while freedom of speech is a fundamental right, it is not

⁴⁰Mchangama, J., Alkiviadou, N., & Mendiratta, R. (n.39) 1-16

⁴¹ Evelyn Mary Aswad, 'The Future of Freedom of Expression Online' (2018) 17 Duke Law & Technology Review 1, pp. 52-53.

⁴²Susan Benesch, 'But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies' (2020) 38 Yale Journal on Regulation Online Bulletin, p.90.

⁴³Voorhoof D (n 20) 1-13

⁴⁴ *Ibid.*

without limitations. The European Court of Human Rights (ECtHR) acknowledges the necessity of safeguarding public order and the rights of others, especially regarding historical events with significant social and moral implications. By establishing these boundaries, the ECtHR ensures a balance where freedom of expression is preserved, but not at the cost of fostering hatred or denying historical facts. This careful approach highlights the crucial role of judicial oversight in interpreting and enforcing free speech protections within the framework of human rights⁴⁵. In the case of *Glimmerveen and Hagenbeek v. the Netherlands*, the ECtHR grappled with the conviction of Nico Glimmerveen and Hans Janmaat, both prominent figures in Dutch far-right politics, for possession of leaflets advocating racial discrimination⁴⁶. The court scrutinised the applicants' intentions, particularly their overarching objective of expelling all non-white individuals from Dutch territory, regardless of humanitarian, economic, or social considerations. The ECtHR's deliberations centred on a meticulous examination of the discriminatory nature of the applicants' expressions, notably their explicit promotion of racial discrimination and the indiscriminate nature of their proposed expulsion policy. Emphasising the potential societal harm stemming from such vitriolic rhetoric, the Court underscored the imperative to safeguard freedom of expression while acknowledging the inherent limitations imposed by hate speech legislation. By prioritising the protection of vulnerable groups and upholding principles of equality and non-discrimination, the ECtHR sought to strike a delicate balance between curtailing expressions that propagate hatred and intolerance while safeguarding the robust exchange of ideas essential to a democratic society. In conclusion, this section has examined the challenges in combating instances of online hate speech, which inherently entails limitations on the freedom of expression. Various actors, including the EU through directives like the DSA, initiatives such as awareness campaigns, and efforts in international law formulation, have sought to mitigate the tension between freedom of expression and the need to address online hate speech. Furthermore, the role of the ECtHR has been crucial in defining boundaries and providing guidance

⁴⁵ *Lehideux and Isorni v. France*, 2000 E.H.R.R.30 665 2000.

⁴⁶ *Glimmerveen and Hagenbeek v. Netherlands*, 4 E.H.R.R. 260, 18 D.R. 187 1982.

in cases concerning hate speech and freedom of expression. Moving forward, the subsequent section will delve deeper into the definition and dynamics of online hate speech, exploring its implications and the evolving responses from legal, societal, and technological perspectives. Next, we delve into the E-commerce Directive, highlighting its ‘safe harbour’ provisions that shield ISPs from liability under specific conditions. The discussion includes the introduction of the DSA, which mandates proactive content moderation by platforms. This section analyses the risks of overreach and the implications for freedom of speech, emphasising the need for balanced regulatory measures. The chapter then provides an overview of the EU and CoE perspectives on online hate speech. It traces the evolution from voluntary codes to binding legislation, critiquing these approaches and evaluating their effectiveness and limitations in managing online hate speech while protecting fundamental rights. We analyse key rulings from the ECtHR and the Court of Justice of the European Union (CJEU) that shape intermediary liability and freedom of expression. Significant cases such as *Delfi AS v. Estonia* and *Google France* are examined to illustrate the legal principles and their application, highlighting the courts' roles in navigating these complex issues. In conclusion, this chapter sets the stage for a comprehensive understanding of the legal landscape surrounding online hate speech. It highlights the ongoing efforts and challenges in balancing regulatory measures with the protection of fundamental rights, providing a foundation for further analysis in subsequent chapters.

2.2 Exceptionalism in regulating online hate speech

The manifestations of hate speech on digital platforms exhibit distinctive characteristics, notably their transnational reach, persistent presence, and the potential for unpredictable resurfacing of content, alongside the perceived anonymity enjoyed by internet users⁴⁷. This phenomenon underscores the imperative for collaborative efforts among nations to prevent and mitigate the grave repercussions stemming from hate speech⁴⁸. However, this endeavour is inherently intricate as it challenges

⁴⁷ Bello, Barbara Giovanna. Tackling online hate speech from a European perspective: Potentials and challenges of inter-legality. Oñati Socio-Legal Series, 2023, pp. 1379-1382.

⁴⁸ *Ibid.*, p. 1387

conventional notions of national jurisdiction and entails divergent perspectives on the classification of content as hate speech within legal frameworks, the interpretation of the bounds of freedom of expression, and the extent of regulatory oversight necessary to foster a secure online environment for all users⁴⁹. These complexities are deeply embedded within cultural and legal contexts, making their resolution highly dependent on specific circumstances⁵⁰. For instance, societies governed by populist movements may face unique challenges in addressing hate speech, as the political climate can significantly influence both the interpretation of what constitutes hate speech and the implementation of regulatory measures⁵¹. In the context of online hate speech, two distinct doctrinal positions can be identified. The first posits that the internet merely transposes content from the physical realm to the virtual domain, implying that additional regulatory measures may be unnecessary or ineffective⁵². This perspective, reflecting a 'non-exceptionalism' stance, advocates for the application of the same legal principles and regulations to the online realm as those governing the offline world⁵³. Conversely, the second doctrine contends that the internet facilitates the proliferation of hate speech through new modes of interaction and content dissemination, thereby necessitating more stringent regulatory and punitive measures⁵⁴. Aligned with an 'exceptionalism' approach towards the web, which posits that the unique characteristics and dynamics of the online environment require distinct

⁴⁹B G Bello (n 49) 1382-1383

⁵⁰*Ibid.*, pp. 1379-1382.

⁵¹ Waldron, J., 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press. See also Ziccardi, G., 2016. *L'odio online: Violenza verbale e ossessioni in rete*. Milan: Raffaello Cortina; Ziccardi, G., 2019. Le espressioni d'odio online. In: G. Ziccardi and P. Perri, eds., *Tecnologia e Diritto*. Milan: Giuffrè Francis Lefebvre, 153–177; Bello, B.G., 2021. I discorsi d'odio in rete. In: T. Casadei and S. Pietropaoli, eds., *Diritto e tecnologie informatiche: Questioni di informatica giuridica. Prospettive, istituzionali e sfide sociali*. Padua: Cedam, 247–262; Bello, B.G., and Scudieri, L., 2022. Discorsi d'odio online. Spunti per un dibattito interdisciplinare. In: B.G. Bello and L. Scudieri, eds., *L'odio online: forme, prevenzione e contrasto*. Turin: Giappichelli, 1–18.

⁵² Ziccardi, G., 2016. *L'odio online: Violenza verbale e ossessioni in rete*. Milan: Raffaello Cortina.

⁵³ Post, D.G., 2008. Governing Cyberspace: Law. *Santa Clara High Technology Law Journal*, 24(4), p.889; see also Fiorinelli, G., 2021. Ordinamenti interconnessi. Il contributo dell'interlegalità alla regolazione della rete. In: E. Chiti, A. di Martino and G. Palombella, eds., *L'era dell'interlegalità*. Bologna: Il Mulino, pp. 405–440.

⁵⁴ Post, D.G., 2008. Governing Cyberspace: Law. *Santa Clara High Technology Law Journal*, 24(4), pp. 889-891.

legal frameworks, we argue that legal frameworks developed for the offline world may not directly apply to online dynamics⁵⁵. Nonetheless, excessive regulation poses the risk of arbitrary censorship, while a lack thereof may exacerbate the proliferation of hate speech⁵⁶. Presently, individuals navigate a complex landscape of legal and regulatory frameworks pertaining to hate speech and internet governance, involving a diverse array of stakeholders ranging from legislators to online platform owners. In their analysis, Pollicino and Bassini⁵⁷ explain how the multitude of regulations affects users, including those propagating or suffering from hate speech, as well as those indirectly affected by interacting with online content. In the case of hate speech content is necessary a rapid removal of harmful content due to the swift dissemination of online hate speech, a task that traditional judicial remedies may not promptly address. Consequently, online intermediaries play a pivotal role in content moderation, albeit at the risk of assuming disproportionate liability and potentially encroaching upon freedom of expression through excessive censorship⁵⁸. Therefore, a viable strategy to address online hate speech, might combine the formulation of appropriate legislative measures by policymakers and the identification and removal of harmful content by online intermediaries. Yet, such an approach necessitates careful consideration of the risks of overregulation to strike a balance between combating hate speech and upholding the principles of freedom of expression.

⁵⁵ Post, D.G., 2008. Governing Cyberspace: Law. *Santa Clara High Technology Law Journal*, 24(4), pp. 889-891.

⁵⁶ Ziccardi, G., (n 52).

⁵⁷ Pollicino, O., and Bassini, M., 2014. The Law of the Internet between Globalisation and Localisation. In: M. Maduro, K. Tuori and S. Sankari, eds., *Transnational Law: Rethinking European Law and Legal Thinking*. Cambridge University Press, pp. 346–380.

⁵⁸ Wu, F.T., 2013. Collateral Censorship and the Limits of Intermediary Immunity. *Notre Dame Law Review*, 87(1), pp. 293–350.

2.3 Balancing Platform Liability and Fundamental Rights in the digital age: The E-commerce Directive

Historically, regulatory efforts pertaining to the liability of intermediaries and platforms have struggled to strike a delicate balance between conflicting rights⁵⁹. This includes the challenge to reconcile freedom of expression with protection from harm, user privacy with content moderation, as well as considerations of innovation and competition versus gatekeeping and control⁶⁰. In the 1990s, initial legislative measures in the EU aimed to establish a balance between user rights and the operational autonomy of Digital Services Providers (DSPs). A fundamental criterion for liability exemptions was grounded on the responsibility of service providers, wherein they minimally intervene in the content flowing through their services, refraining from filtering or deleting it⁶¹. This principle was enshrined in the E-Commerce Directive, a pivotal moment in delineating the accountability of online platforms within the EU⁶². The differing methods of the CJEU, which promotes a vibrant online market, and the ECtHR, which focuses on protecting fundamental human rights and freedom of expression, highlight the distinct ways each addresses the responsibilities of online intermediaries while ensuring the protection of fundamental rights. Consequently, the EU legislator decided that Digital Services Providers, encompassing both hosting providers and access providers, should benefit from immunity from liability for the actions of third parties on their

⁵⁹ C. Angelopoulos and S. Smeth, 'Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability', 2017, 8(2) *Journal of Media Law*, 1–26 <https://www.ivir.nl/publicaties/download/Notice_and_Fair_Balance.pdf>.

⁶⁰ Maroni, Marta, and Elda Brogi. "Freedom of expression and the rule of law: the debate in the context of online platform regulation", *Research handbook on EU media law and policy*. Edward Elgar Publishing, 2021, pp. 166-189; See Sander, Barrie. "Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation." *Fordham Int'l LJ* 43 (2019), pp.939-1006.

⁶¹ Passive Intermediaries (Mere Conduit and Caching): These intermediaries act as neutral channels for information. They don't modify or select the content and have limited knowledge of it. Examples include: Internet Service Providers (ISPs), Domain name registrars (Article 12 and Article 13 of Directive 2000/31/EC). Active Intermediaries (Hosting): These intermediaries play a more substantial role. They store information and may even organise or promote it. Examples include: Social Media Platforms and Online Marketplaces (Article 14 Directive 2000/31/EC).

⁶² Frosio G and Geiger C (n 9) 31-77.

networks⁶³. The EU influenced by the US legislator's⁶⁴ lead with the Communications Decency Act and the Digital Millennium Copyright Act, swiftly adopted similar provisions in the E-Commerce Directive, mandating Member States (MSs) to implement safeguarding DSPs from liability under certain conditions⁶⁵. The distinction between passive and active hosts stems from the expansive interpretation of Recital (42) in the E-Commerce, which mandates that intermediaries' functions remain strictly technical, automatic, and passive. The distinction between passive and active hosts stems from the expansive interpretation of Recital (42) in the E-Commerce, which mandates that intermediaries' functions remain strictly technical, automatic, and passive⁶⁶. The Court of Justice of the European Union (CJEU) has played a significant role in clarifying this distinction through landmark cases such as *Google France*⁶⁷ and *L'Oréal v. eBay*⁶⁸. In these cases, the CJEU established that a passive host is one that merely provides the infrastructure for storing or transmitting information without engaging in any active role that could give it knowledge or control over the content. Passive hosts perform functions that are purely technical, automatic, and neutral, and they do not have the ability to modify the information they handle. Conversely, an active host is one that goes beyond these passive functions and takes an active role in managing, organizing, or promoting the content. This involvement can give the host knowledge or control over the data, thus making it more liable for the content. Thus, the classification between active and passive hosts is crucial as it determines

⁶³ Wilman, Folkert. "The EU's System of Knowledge-Based Liability for Hosting Services Providers in Respect of Illegal User Content-Between the e-Commerce Directive and the Digital Services Act." *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12 (2021), pp. 318-321.

⁶⁴ See Communications Decency Act of 1996, 47 U.S.C. § 230; The Digital Millennium Copyright Act of 1998, 17 U.S.C. § 512.

⁶⁵ 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 [2000] OJ L 178/1–16.

⁶⁶ Recital 42, 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'), Official Journal L 178 , 17/07/2000 P. 0001 - 0016.

⁶⁷ *Google France SARL and Google Inc. v Louis Vuitton Malletier SA* (Joined Cases C-236/08 to C-238/08), Judgment of the Court (Grand Chamber) of 23 March 2010.

⁶⁸ *L'Oréal SA and Others v. eBay International AG and Others* (Case C-324/09), Judgment of the Court (Fourth Chamber) of 12 July 2011.

the extent of liability that intermediaries may face for the content they host. Passive hosts benefit from broader liability exemptions due to their limited role in handling the content, whereas active hosts, due to their enhanced engagement, bear greater responsibility and potential liability for the content on their platforms. This distinction ensures that while intermediary service providers can facilitate the free flow of information, they are also held accountable when they take on roles that influence or control the content, thereby impacting the legal and regulatory landscape of digital services. Regarding the attribution of responsibility for illicit content, a cornerstone of the regulatory framework devised by the EU lies in Articles 12 to 15 of the Electronic Commerce Directive (ECD⁶⁹), specifically addressing the liability of intermediary service providers⁷⁰. The legislation delineates three distinct categories of liability which depend on the role undertaken by the providers: 'mere conduit,' 'caching,' and 'hosting,'. Providers classified within the first two categories ('mere conduit' and 'caching') under the E-Commerce Directive are granted immunity from liability, akin to the broad protections provided by Section 230(c)(1) of the US Communications Decency Act⁷¹. However, the conditions and scope of immunity differ, with the E-Commerce Directive imposing specific requirements based on the nature of the intermediary's role. 'Mere conduit' service providers are exempted from liability when the service provider is only passively involved in the transmission of data (i.e. interconnection to the internet provided by traditional internet service providers and network operators)⁷². Caching providers are not liable when they temporarily and automatically store data to improve transmission efficiency (such as with proxy servers), provided they meet certain technical conditions, like ensuring the local copy matches the original. Service providers classified as 'hosting'

⁶⁹ 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.

⁷⁰ *Ibid.*

⁷¹ Communications Decency Act (CDA), Pub. L. No. 104-104 (Tit. V), 110 Stat. 133 (Feb. 8, 1996).

⁷² Article 12, 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.

are absolved from liability for stored information if two conditions are satisfied: (a) the Internet Service Providers (ISPs) lacks actual knowledge of illegal activities or, in the case of claims for damages, lacks awareness of facts or circumstances from which illegal activity is apparent; or (b) upon obtaining such knowledge or awareness, the provider acts expeditiously to remove or disable access to the information⁷³. The ‘hosting’ providers are exempted from liability when those companies’ storing data for their users (e.g. webhosting) do not know that they host illegal activity or information and act expeditiously to remove or disable access to the illegal information⁷⁴. Any actions undertaken by service providers must adhere to Article 10 of the European Convention on Human Rights (ECHR)⁷⁵, which safeguards the right to freedom of expression for all individuals. This right includes the freedom to hold opinions, and to receive and impart information and ideas without interference by public authority and regardless of frontiers. Article 10 stipulates that the exercise of these rights may be subject to certain restrictions, conditions, or formalities. However, any such restrictions must withstand scrutiny under a three-pronged test to ensure non-arbitrary actions. Any intervention must (1) be prescribed by law, (2) pursue a legitimate aim, and (3) be necessary in a democratic society. This necessity test involves assessing whether the restriction is (a) appropriate for achieving the legitimate aim (fitness), (b) the least restrictive means available (necessity), and (c) proportionate to the aim pursued (*stricto sensu* proportionality)⁷⁶. While the ECtHR exercises a balancing assessment between the right to free speech and the online hate speech material, it remains critical to determine whether the content removal is legal or amounts to

⁷³Article 14, 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 1-16.

⁷⁴*Ibid.* See also Tambiama Madiaga, "Reform of the EU liability regime for online intermediaries," European Parliamentary Research Service, PE 649.404 – May 2020.

⁷⁵ Council of Europe. (1950). *European Convention on Human Rights*. Article 10.

⁷⁶Jasper P. Sluijs, From Competition to Freedom of Expression: Introducing Article 10 ECHR in the European Network Neutrality Debate, *Human Rights Law Review*, Volume 12, Issue 3, September 2012, Pages 509–554, <https://doi.org/10.1093/hrlr/ngs015>

censorship⁷⁷. The initial schemes of liability exemption aimed to balance the prohibition of harmful content with the preservation of freedom of expression⁷⁸. These schemes were designed with the premise that safe harbours would incentivise Digital Services Providers to prevent the hosting of illegal content on their platforms, thereby promoting public interests such as protecting children, combating hate speech, and safeguarding consumer rights⁷⁹. Under Article 14(1) of the E-Commerce directive, DSPs are granted a safe harbour from liability, on the condition that they lack awareness of the infringement and promptly take action to remove or restrict access to the infringing content upon knowledge of its presence on their network⁸⁰. Conversely, Article 15(1) prohibits member countries from imposing a broad obligation on Internet Service Providers (ISPs) to monitor or control the storage or transmission of information, or to request evidence of suspected illegal activities, as stipulated in Articles 12-14⁸¹. The E-Commerce Directive establishes a system of safe harbor provisions where intermediaries are exempt from liability under certain conditions, such as not having actual knowledge of illegal content and acting expeditiously to remove or disable access to such content once they become aware of it. The Directive also prohibits Member States from imposing a general obligation on intermediaries to monitor the information they transmit or store⁸². By functioning retrospectively rather than prospectively, the liability exemptions for 'hosting' providers, as outlined in the Directive, promote the policy goal of reducing 'chilling effects.' This is particularly important given the significant impact of viral and widely distributed content. The 'chilling effect' refers to the inhibition or restraint of speech or expression prompted by apprehension about potential

⁷⁷Gergely Gosztonyi, How the European Court of Human Rights Contributed to Understanding Liability Issues of Internet Service Providers, *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae Sectio Iuridica* (2019).

⁷⁸Geiger, Christophe, Giancarlo Frosio, and Elena Izyumenko. "Intermediary liability and fundamental rights." *The Oxford Handbook of Intermediary Liability Online* (OUP, 2020) (2019).

⁷⁹*Ibid.*

⁸⁰ Frosio G and Geiger C (n 9) 31-77.

⁸¹ E-Commerce Directive (ECD), "Article 15 No general obligation to monitor", 2000.

⁸² Frosio G and Geiger C (n 9) 31-77.

legal repercussions or societal condemnation⁸³. In essence, the E-Commerce Directive offers legal protections to access and hosting providers regarding illicit activities carried out by users on their platforms, covering a range of violations. The directive establishes a liability framework based on two fundamental principles: the principle of actual knowledge and control, and the prohibition of general monitoring obligations. These principles aim to balance the facilitation of user accessibility with the protection of their rights. The principle of acquisition and knowledge stipulates that ISPs are liable for hate speech content on their platforms only when they are aware of its presence, requiring prompt action upon notification⁸⁴. Conversely, the non-monitoring obligation principle exempts ISPs from actively monitoring content, advocating a reactive approach to intervention after specific notifications of hate speech, thus mitigating surveillance burdens and preserving user privacy⁸⁵. These principles establish a legal framework that balances combating hate speech with protecting ISPs' rights and promoting equitable content moderation practices online. If there are incidents that relate to online hate speech, the E-Commerce Directive does not grant complete immunity to ISPs. While an 'access' provider is largely protected if it only plays a passive role in transmitting content, the situation is different for a 'hosting' provider⁸⁶. The E-Commerce Directive offers a partial solution to liability in relation to Intermediary Service Providers and Hosting Service Providers, falling short of being a complete remedy. The system of conditional liability exemption provided by the Directive creates a significant financial incentive for private censorship. Essentially, hosting providers are motivated to remove any content flagged as illegal, regardless of its actual legality, to safeguard

⁸³Schauer, Frederick. "Fear, risk and the first amendment: Unravelling the chilling effect." *BUL rev.* 58 (1978), pp. 685-732

⁸⁴ Article 14 of the E-Commerce Directive provides that ISPs (specifically hosting providers) are exempt from liability for user-generated content unless they have actual knowledge of illegal activity. Upon obtaining such knowledge, they must act expeditiously to remove or disable access to the content to maintain their safe harbor protection.

⁸⁵Article 15 of the E-Commerce Directive explicitly states that Member States shall not impose a general obligation on ISPs to monitor the information they transmit or store, nor to actively seek facts or circumstances indicating illegal activity. This principle helps mitigate the burden of surveillance on ISPs and protects user privacy.

⁸⁶ Isabelle Rorive, What Can Be Done against Cyber Hate - Freedom of Speech versus Hate Speech in the Council of Europe, 17 *CARDOZO J. INT'L & COMP. L.* 417, 2009, pp. 417-426

themselves from potential liability. Adhering to the Directive allows these providers to benefit from provisions that shield them from non-contractual liability. Often, hosting providers include clauses in their service terms and conditions to protect themselves against any claims for damages arising from unlawful content posted by their customers⁸⁷. The focal contention underscores the critical dichotomy between the interests of private entities and platforms and those of the public sphere concerning the structuring of liability. However, there is a prevalent scenario where states exert pressure on platforms to comply with specific standards, ostensibly framed as safeguarding citizens' interests, yet occasionally serving to advance state-centric agendas and curtail freedom of expression⁸⁸.

2.4 Is there a definition of on online hate speech? The CoE legislative framework.

The examination of online hate phenomena poses multifaceted challenges across political, legal, regulatory, and technological domains concerning the governance and oversight of the virtual landscape of the internet⁸⁹. As already stated in previous sections, this intricate process involves diverse stakeholders, including online service providers⁹⁰. The threats associated with hate speech have been acknowledged by the European Union, as evidenced in its analysis of the E-Commerce

⁸⁷ For an overview of this issue, see Isabelle Rorive, *Strategies to Tackle Racism and Xenophobia on the Internet-Where Are We in Europe?* 7 INT'L J. COMM. L. & POL'Y 8 (2003), available at http://www.ijclp.org/7_2003/pdf/rorive-ijclp-artikel.pdf; Isabelle Rorive, "Regulating the Internet?", address at the Conference on the Council of Europe's Convention on Information and Legal Co-operation Concerning "Information Society Services," May 21, 2003, available at [http://www.coe.int/T/E/Legal-Affairs/Legal-co-operation/Information Society-Services/Riguler%201%27Internet%20-%20CoE%20%2021%20%20Mai%202003%20_version%20icrite_.pdf](http://www.coe.int/T/E/Legal-Affairs/Legal-co-operation/Information_Society-Services/Riguler%201%27Internet%20-%20CoE%20%2021%20%20Mai%202003%20_version%20icrite_.pdf).

⁸⁸ Frosio and Geiger (n 9) 31-77.

⁸⁹ Bassini, M., 2019. Fundamental Rights and Private Enforcement in the Digital Age. *European Law Journal*, 25(2), 182–197; See also Ziccardi, G., and Perri, P., 2022. L'odio online tra profilazione, big data e protezione dei dati personali. In: B.G. Bello and L. Scudieri, eds., *L'odio online: forme, prevenzione e contrasto*. Turin: Giappichelli, 91–106.

⁹⁰ Kettemann, Matthias C. The normative order of the internet: A theory of rule and regulation online. Oxford University Press, 2020, pp. 8-25. See also Fiorinelli, G., 2021. Ordinamenti interconnessi. Il contributo dell'interlegalità alla regolazione della rete. In: E. Chiti, A. di Martino and G. Palombella, eds., *L'era dell'interlegalità*. Bologna: Il Mulino, 405–440.

Directive discussion earlier. While there is a lack of a universally accepted legal definition of hate speech, reflections on what constitutes hate speech persist. There is also an overarching consensus that severe infringements on freedom of expression posing threats to democratic principles and social harmony necessitate robust intervention⁹¹. A definition of hate speech within the context of the CoE can be inferred from the Additional Protocol to the Convention on Cybercrime, which primarily addresses manifestations of racial or xenophobic animosity. This protocol focuses on criminalising acts such as the dissemination of racist and xenophobic material through computer systems⁹². Cyber-hate is described as the utilisation of inflammatory or derogatory language online targeting attributes such as religion, race, gender, or political affiliations⁹³. The Council of Europe's Convention on Cybercrime, also known as the Budapest Convention, ratified in 2001, serves as the principal legally binding instrument internationally for regulating criminal activities conducted on the Internet. Its objective is to standardize national criminal laws concerning cybercrime by providing guidelines for national investigative and prosecutorial processes and establishing a robust framework for international cooperation⁹⁴. Article 23 of the European Cybercrime Treaty emphasises the principles of international co-operation, urging states to exert maximum effort in investigating and prosecuting computer-related crimes and gathering electronic evidence using both international instruments and relevant national laws⁹⁵. However, it does not offer specific instructions for harmonising national legislations. Countries that have ratified the 2006 Additional Protocol to the Convention on

⁹¹Alkiviadou, Natalie. "The legal regulation of hate speech: The international and European frameworks." *Politická misao*, 2018, pp. 203-229.

⁹²Council of Europe. Additional protocol to the convention on cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems: Strasbourg, 28. 1. Council of Europe Publ., 2003.

⁹³ Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, Harvey Mauricio Herrera López, Internet, social media and online hate speech. Systematic review, *Aggression and Violent Behaviour*, Volume 58, 2021, 101608, ISSN 1359-1789, <https://doi.org/10.1016/j.avb.2021.101608>(<https://www.sciencedirect.com/science/article/abs/pii/S1359178921000628#ab0010>), pp.1-7.

⁹⁴ Council of Europe Convention on Cybercrime. 23 November 2001. *European Treaty Series*, No. 185.

⁹⁵ *Ibid.*

Cybercrime, which criminalises acts of a racist and xenophobic nature committed via cyber means must comply with it⁹⁶. In addition to the Treaty, several non-binding measures, such as the Recommendation on Combating Hate Speech adopted in 2022, aim to address such speech both online and offline while upholding human rights, democracy, and the rule of law⁹⁷. While it advocates for a unified definition of hate speech⁹⁸ to encourage alignment of legal definitions across governments, it defers the determination of which expressions fall outside the purview of freedom of speech to individual states. It mandates that restrictions on freedom of expression adhere to the ECHR and relevant ECtHR case law, acknowledging potential variations in court rulings. Furthermore, it advocates for a comprehensive approach involving diverse stakeholders, including public officials, media, civil society organizations, and internet intermediaries. The Council of Europe Recommendation CM/Rec (2022) emphasises that national regulations concerning online hate speech should adhere to established standards outlined in the Recommendation CM/Rec (2018)2 on the Roles and Responsibilities of Internet Intermediaries⁹⁹. The Appendix provides further guidance on national provisions pertaining to internet intermediaries' jurisdiction, reporting mechanisms, removal procedures, and the role of judicial review in addressing online hate speech¹⁰⁰.

2.5 Evolution and Critique of European Union Regulatory Approaches to Online Hate Speech: From Voluntary Codes (Soft Law) to Hard-Law Measures

Proceeding with our analysis, this section explores how the European Union's regulatory and self-regulatory approaches to tackling online hate speech have changed over time. We begin by analysing

⁹⁶Council of Europe Additional Protocol to the Convention on Cybercrime, concerning the Criminalisation of Acts of a Racist and Xenophobic Nature Committed through Computer Systems, Strasbourg, 28.1. *European Treaty Series*, No. 189, 2003.

⁹⁷ Council of Europe Recommendation CM/Rec (2022)16 of the Committee of Ministers to member States on combating hate speech, para.5 (Adopted by the Committee of Ministers on 20 May 2022 at the 132nd Session of the Committee of Ministers).

⁹⁸*Ibid.*, para.17.

⁹⁹*Ibid.*, para.17.

¹⁰⁰*Ibid.*, paras.18–27.

the impact of less binding initiatives, such as the Codes of Conduct adopted by the EU. This analysis is instrumental in understanding the excursus that led the EU to adopt the DSA of 2022 and allows us to better comprehend the consequences of these measures have on internet governance and the protection of fundamental rights, especially those of internet users. Over time, the EU has launched initiatives ranging from informal agreements on codes of conduct with major online platforms to the adoption of binding legislation¹⁰¹. The aim of these initiatives has been to harmonise the legislative provisions across Member States (MSs) on various aspects of internet governance, focusing on relations with online intermediaries. Regarding informal EU agreements, in 2016 the European Commission reached an accord with Facebook, Microsoft, Twitter and YouTube on a Code of Conduct aimed at countering illegal online hate speech¹⁰². Subsequently, in 2018, Instagram, Snapchat and Dailymotion joined the initiative, followed in 2019 by Jeuxvideo.com and in spring 2022 by Rakuten, Viber and Twitch. The need to adopt the ‘EU Code of Conduct on Countering Illegal Hate Speech Online’ arose due to differences between the US and European approaches to protection against hate speech and freedom of expression. The Code introduced stricter rules that deviated from the North American legal culture of the major platforms. In terms of enforcement, these platforms committed to prohibiting the promotion of behaviour inciting violence and hatred, adopting timely and effective processes to handle reports of such phenomena, assessing these reports in the light of EU legislation, including, where necessary, the national laws of the Member States implementing the Framework Decision¹⁰³, and removing illegal content in a timely manner. Regarding prevention, platforms are committed to raising awareness of hate content among their users. The implementation of the Code of Conduct is regularly monitored through an agreed methodology based on checking the response rates, transparency analysis, and evaluation of content

¹⁰¹European Commission: The EU Code of conduct on countering illegal hate speech online. Available at: <https://www.coe.int/en/web/cyberviolence/-/european-commission-the-eu-code-of-conduct-on-countering-illegal-hate-speech-online>

¹⁰²*Ibid.*

¹⁰³Council Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 28 November 2008 (2008/913/JHA).

removal decisions in cooperation with a network of organisations in several EU countries. The main provisions of the Code of Conduct are: (a) Platforms must provide clear, publicly available, clearly guidelines stipulating what constitutes hate speech or harmful content. (b) Platforms must have a system to review user notifications of hate speech content. (c) Platforms must remove illegal content upon verification of its illegality. (d) Platforms are required to provide regular reports on the number of notifications received and actions taken. However, monitoring reports have show uneven progress over time, with a decrease in the rates of assessment and removal of illegal hate content¹⁰⁴. The 2016 Code was evaluated, revealing several limitations in relation to its effectiveness¹⁰⁵. First, its voluntary and non-binding nature limits its ability to ensure continuous compliance. MSs and other actors like European Commission, online platforms, civil society organisations, and relevant committees responsible for monitoring the Code are not required to comply with the Code as it is a form of soft law. Periodic evaluations, primarily quantitative in nature, emphasise the number of illegal hate speech instances removed by platforms rather than the quality or thoroughness of the decision-making process¹⁰⁶. The quantitative approach emphasises the quantity of number of instances of illegal hate speech content removed by the online platforms, rather than the quality or thoroughness of the decision-making process behind these removals¹⁰⁷. Moreover, quantitative procedures do not adequately address the legality of content nor provide transparency on the use of resources for content investigation. Additionally, the Code raises concerns about a lack of oversight and accountability, as it allows IT companies to exercise significant authority, resembling legislative, judicial, and executive

¹⁰⁴ Reynders, Demperti., *6th Evaluation of the Code of Conduct. Factsheet* [online]. European Commission, Directorate-General for Justice and Consumers, 7 October. 2021, pp. 99. Available from: https://ec.europa.eu/commission/presscorner/detail/en/fs_21_5106 [Access 10 March 2024].

¹⁰⁵ European Commission Factsheet, 'Code of Conduct on Countering Illegal Hate Speech Online: First Results on Implementation', at 1 (Dec. 2016), <http://ec.europa.eu/information society/newsroom/image/document/2016-50/factsheet-code-conduct-840573.pdf>

¹⁰⁶ *Ibid.*

¹⁰⁷ *Ibid.*

powers¹⁰⁸. For instance, internal guidelines of individual companies, which define hate speech, sometimes override national regulations or the Code's own definition of hate speech¹⁰⁹. This inconsistency raises concerns regarding the definition and application of hate speech rules. Moreover, the lack of legal expertise of the content moderators of digital platforms, positions them as arbiters of fundamental rights without the necessary training or judicial expertise. Another criticism is that the absence of EU legislation on protecting vulnerable groups leaves IT companies responsible for setting fundamental rights rules without a legislative mandate. This situation raises questions about national sovereignty and the consistent application of hate speech rules across the EU¹¹⁰. Despite these criticisms, the Code has practical advantages such as flexibility, efficiency and usefulness in cross-border enforcement. Its adaptable nature encourages participation from stakeholders without imposing excessive regulation and allows IT companies to effectively manage reported content using their significant financial resources. Employees are more familiar with the platforms, and the considerable financial resources of companies allow them to better deal with the management of reported content¹¹¹. Another argument in favour of the Code is that it offers consistent enforcement across Europe, mitigating national disparities in the online hate speech regulatory system¹¹². However, this work argues that to improve the Code's effectiveness, it might be useful to include qualitative analysis such as thematic and content analysis as part of periodic reviews, thus ensuring that legal content is not wrongly removed. Such qualitative analysis methods can help to understand themes and contents of published material. It is important to note that the 2016 Code of Conduct remains in force and operates alongside hard law measures such as the Digital Services Act (DSA),

¹⁰⁸Quintel, T., & Ullrich, C. (2018). Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond. In B. Petkova & T. Ojanen (Eds.), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries* (pp. 1-21). Edward Elgar Publishing.

¹⁰⁹ 'The Code of Conduct on Countering Illegal Hate Speech Online' (EU Commission, May 2016).

¹¹⁰Quintel T and Ullrich C (n 109) 123-125.

¹¹¹ *Ibid.*, pp.123-125

¹¹²James Banks, 'European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation' *European Journal of Crime, Criminal Law and Criminal Justice*, 19 (1), 1-13, 2011.

creating a dual framework of both soft and hard law to address online hate speech. The Enhanced Disinformation Code of Practice 2022 further updates the self-regulatory standards for combating disinformation introduced in the 2018 Code¹¹³. The Strengthened Code of Practice on Disinformation 2022 is another soft law measure based on the voluntary approach of exchanging good practices. It is a voluntary agreement between the European Commission and online platforms¹¹⁴. This new Code, signed by platforms such as Meta, Vimeo, Twitter and TikTok, and presented in June 2022, includes measures to increase algorithmic transparency, demonetise misinformation content, and enhance cooperation with researchers. Meanwhile, the EU has gradually shifted towards a hard-law approach, exemplified by Directive 2018/1808¹¹⁵, which amends the Audiovisual Media Services Directive (AVMSD). The directive aims to ensure the effectiveness of self- and co-regulatory measures to protect consumers and public health, particularly in the context of advertising, hate speech, and harmful content on audiovisual media and video-sharing platforms¹¹⁶. This Directive introduces enforceable rules to protect minors and users from harmful audio-visual content and hate speech shared online¹¹⁷, thus strengthening the protection against hate speech. Member States are also invited to provide for out-of-court redress mechanisms to resolve disputes between users and providers of video-sharing platforms. Yet, there was no specific time frame for this mechanism, which could cause laxity in the provision of solutions to resolve online disputes. The Directive 2018/1808, known as the Audiovisual Media Services Directive (AVMSD), was updated in 2018 in view of significant changes in society's media consumption patterns, with a marked shift towards the use of video-sharing

¹¹³ European Commission, 2022 Strengthened Code of Practice on Disinformation [online]. 16 June. Brussels. Available from: <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>.

¹¹⁴ Monti, Matteo. "The EU Code of Practice on Disinformation and the Risk of the Privatisation of Censorship." *Democracy and Fake News*. Routledge, 2020, pp. 214-225.

¹¹⁵ European Parliament and Council Directive 2018/1808 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, 14 November 2018 (PE/33/2018/REV/1). *Official Journal*, L 303, 28.11.2018.

¹¹⁶ *Ibid.*

¹¹⁷ *Ibid.*

platforms¹¹⁸. This shift has been accompanied by growing concerns regarding 'harmful content' and 'hate speech' present on such platforms¹¹⁹. Consequently, the AMSD extended its scope to include these platforms (YouTube, TikTok, Vimeo, and Dailymotion) alongside traditional television broadcasting and video-on-demand services, placing the onus on platforms to take measures to protect minors and users from 'harmful content' such as hate speech¹²⁰. The directive emphasises the need to carefully balance these measures with fundamental rights, including freedom of expression¹²¹. This expansion of the directive has brought some benefits, such as bringing both new (online) and old (broadcast) media under a more cohesive regulatory framework, thus potentially simplifying the supervision process¹²². It also introduced enforceable rights that aim to make the Internet a safer space, especially for minors¹²³. However, these efforts are not without criticism for three main reasons. Firstly, the definitions of 'harmful content' for minors and 'hate speech' for the public are not sufficiently specific, leaving national legislations too much room for interpretation. Secondly, the

¹¹⁸ EU Commission Fact Sheet on AMSD: Children are watching less traditional TV: the average daily viewing time for young Europeans was 2 hours in 2014 i.e. about half as much as the average viewer. Furthermore, global internet video share in consumer internet traffic is expected to increase from 64% in 2014 to 80% by 2019.

¹¹⁹ European Parliament and Council Directive 2018/1808 of 14th 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 (2018) OJ L 303/69 (from herein referred to as the 'AMSD'), recital 45.

¹²⁰ European Parliament and Council Directive 2018/1808 of 14th 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 (2018) OJ L 303/69 (from herein referred to as the 'AMSD'), article 28(b).

¹²¹ *Ibid.*, recital 51.

¹²² European Parliament and Council Directive 2018/1808 of 14th 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 (2018) OJ L 303/69 (from herein referred to as the 'AMSD'), Article 30.

¹²³ *Ibid.*, articles 6, 6(a); It should be noted that alongside these enforceable rights are the 'Terms of Service' legal provisions on hate speech as found on all major social media platforms. It has been noted elsewhere that said 'ToS' provisions – although unregulated – often go beyond those provided at national and international level. See: Joan Barata 'Regulating content moderation in Europe beyond the AVMSD' (Media@LSE Blog, February 25th, 2020).

way the 'country of origin' jurisdictional principle of the AMSD is applied remains uncertain, especially considering that many of the major platforms are based in Ireland¹²⁴. Ireland is often perceived as having relatively relaxed content moderation practices due to its regulatory framework and the presence of many major tech companies' European headquarters in the country. This means content that might be deemed harmful in other EU countries could potentially slip through the cracks under Irish regulations. Consequently, the Irish regulator plays a pivotal role in defining and moderating sensitive and controversial content at the European level, despite the residence of the content's author and its target audience¹²⁵. Thirdly, expanding the liability of intermediaries by updating the obligations of platforms may violate Article 15 of the E-Commerce Directive, which prohibits general monitoring obligations. This could effectively represent a revision of the liability framework for Internet intermediaries, raising concerns about compliance with established EU law¹²⁶. This raises concerns among human rights advocates who fear that an increased threat of liability may push platforms to pre-emptively remove any potentially controversial material.

2.6 The Jurisprudential Dynamics of Hate Speech Liability in Online Platforms within the European Court of Human Rights (ECHR) in Strasbourg and the Court of Justice of the European Union (CJEU)

Analysing the legal precedents set by prominent European tribunals regarding the nuanced assignment of liability for online service providers and the application of the ECHR, particularly by the ECtHR in Strasbourg and the CJEU in Luxembourg, helps understand the broader European

¹²⁴Lucas Volman, L. L. "Is the cornerstone loose? Critical analysis of the functioning of the 'country of origin's principle in the Audiovisual Media Services Directive, taking into account the rapid changes in the audiovisual industry and the recent challenges brought by Brexit." 2018.

¹²⁵*Ibid.*

¹²⁶ European Parliament and Council Directive 2018/1808 of 14th 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 (2018) OJ L 303/69 (from herein referred to as the 'AMSD'), article 28(b)(3): "...measures shall not lead to any ex-ante control measures or upload-filtering of content which do not comply with Article 15 of Directive 2000/31/EC."

approach to holding ISPs accountable in upholding freedom of expression and addressing hate speech-related censorship. A fundamental distinction exists between the roles played by the CJEU and the ECtHR. The CJEU acts as a catalyst for greater cohesion within the EU, while the ECtHR ensures a fundamental level of protection for human rights standards¹²⁷. The CJEU adjudicates conflicts involving EU institutions or between member states, while also ensuring that member states and EU institutions comply with the EU treaties and laws domestically¹²⁸. Additionally, since the Maastricht Treaty, the CJEU has the authority to impose penalties on legal entities and member states found to have breached EU legislation. This authority is outlined in Articles 258-260 of the Treaty on the Functioning of the European Union (TFEU)¹²⁹. Conversely judgments issued by the ECtHR are likely to result in slower legislative changes with potentially lower political implications compared to similar rulings by the CJEU due to the different nature of the two judiciaries¹³⁰. These two judicial bodies consider each other's judgments regarding human rights infringements. The interaction between the two Courts is significant; while the ECtHR serves as a constitutional court emphasising fundamental rights, the CJEU has been notably influenced by the original economic objectives of the European Community¹³¹. These differing approaches are evident in how they handle

¹²⁷ Gosztonyi, G., How the European Court of Human Rights Contributed to Understanding Liability Issues of Internet Service Providers. *Annales Universitatis Scientiarum Budapestinensis De Rolando Eötvös Nominatae. Sectio Iuridica*, 58, 2019, pp. 121–133. <https://doi.org/10.56749/Annales.elteajk.2019.lviii.7.121>

¹²⁸ S. Bardarova, Comparison Between the European Court of Justice and European Court of Human Rights, SSRN Paper, (18 June 2013), <https://doi.org/10.2139/ssrn.2281215>, pp.1-12.

¹²⁹ Articles 258-260 TFEU describe the infringement procedure and the CJEU's role in enforcing EU law and imposing penalties. Available at: <http://data.europa.eu/eli/treaty/tfeu/2012/oj>. For an in-depth comparison of the roles of the CJEU and the ECtHR, see S. Bardarova, Comparison Between the European Court of Justice and European Court of Human Rights, SSRN Paper, (18 June 2013), <https://doi.org/10.2139/ssrn.2281215>.

¹³⁰ For further discussion on the nature of the two Courts and their role in relation to fundamental rights see: S. Morano-Foadi and S. Andreadakis 'Protection of Fundamental Rights in Europe: The Challenge of Integration', Springer, June 2020; S. Morano-Foadi and L. Vickers A matter for two courts: the fundamental rights question for the EU, Hart Publishing, 2015; E. Imbarlina, The Roles and Relationship between the Two European Courts in Post-Lisbon EU Human Rights Protection, *Jurist*, (12 September 2013) <https://www.jurist.org/commentary/2013/09/elena-butti-lisbon-treaty>

¹³¹ O. Pollicino, Judicial protection of fundamental rights in the transition from the world of atoms to the word of bits: the case of freedom of speech, (2019) 25 (2) *European Law Journal*, 168. <https://doi.org/10.1111/eulj.12311>. See also 'S. Morano-Foadi and S. Andreadakis Protection of Fundamental

matters related to content regulation and moderation by platform providers. The contrasting approaches of the CJEU, which emphasises fostering a dynamic online marketplace, and the ECtHR, which prioritises the safeguarding of fundamental human rights, particularly freedom of expression, underscore the different methodologies employed by each in addressing the obligations of online intermediaries while ensuring the protection of fundamental rights. Therefore, it is essential to examine how the ECtHR addresses these concerns, as it, alongside the CJEU, has played a substantial role in shaping recent digital regulation initiatives, notably the Digital Services Act. This suggests a trend towards the CJEU aligning with contemporary liability demands. The ECtHR has shed light on potential public interest concerns arising from the activities of digital service providers (DSPs), such as the right to privacy versus data collection and freedom of expression versus harmful content¹³². The extent of permissible constraints on freedom of expression varies depending on the specific content under scrutiny. For instance, in *Handyside v. United Kingdom* (1976)¹³³, the ECtHR ruled that the book's content, while potentially offensive to some, did not warrant such a severe restriction on freedom of expression. This case is a landmark decision by the ECtHR concerning the balance between freedom of expression and public morality. Richard Handyside, a British publisher, was convicted under the UK Obscene Publications Act for distributing "The Little Red Schoolbook," a book intended to educate children about sex, drugs, and other social issues, which was deemed obscene by the UK authorities. Handyside argued that his conviction violated Article 10 of the European Convention on Human Rights, which protects freedom of expression. The ECtHR ruled that although freedom of expression is fundamental, it is subject to certain restrictions necessary in a

Rights in Europe: The Challenge of Integration', Springer, June 2020; S.Morano-Foadi and L. Vickers - A matter for two courts: the fundamental rights question for the EU, Hart Publishing, 2015. 'Reflections on the Architecture of the EU after the Treaty of Lisbon: the European judicial approach to fundamental rights' in Special Issue edited by S Morano-Foadi and F Duina, entitled "'The Institutionalisation of Regional Trade Agreements Worldwide: New Dynamics and Future Perspectives', *European Law Journal*, Vol. 17, No. 5, September 2011, 595–610.

¹³² See *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015); *Satakunnan Markkinapörssi Oy and Satamedia Oy v. Finland* (Application No. 931/13).

¹³³ *Handyside v. United Kingdom* (1976) (Application no. 5493/72).

democratic society. The Court upheld the UK's margin of appreciation, allowing national authorities discretion in determining what is obscene, given their proximity to the cultural and moral values of their society. This case is significant because it established the principle of margin of appreciation, whereby national authorities are given some leeway in balancing individual rights and public interests, recognising the diversity of moral values across Europe. It also underscored that while freedom of expression is crucial, it can be lawfully restricted to protect public morality, provided such restrictions are proportionate and necessary in a democratic society. Courts have employed diverse evaluative criteria, considering the inherent characteristics of the content subject to moderation¹³⁴. Furthermore, the ECtHR has emphasised the importance of addressing or curbing expressions that propagate hatred stemming from intolerance, provided that any limitations imposed align with legitimate aims pursued in a proportional manner¹³⁵. For example, in *Verein Gegen Rechtsmissbrauch (VGR) v. Germany (2018)*, the Court upheld a German court's decision to block a neo-Nazi website filled with hate speech and incitement to violence. Certain types of expression, such as denial of crimes against humanity and well-established historical truths, are ineligible for protection under Article 10 of the ECHR, as outlined in Article 17, which prohibits the abuse of rights¹³⁶. Hence, such expressions should be swiftly removed with appropriate constraints. This rigorous stance was reaffirmed in the case of *Delfi AS v. Estonia*, which pertained to the delayed removal of online remarks containing hate speech and calls to violence. The Court emphasised the responsibility of platforms to monitor and promptly address harmful content to protect individuals from hate speech, ensuring that any limitations on freedom of expression are proportionate and necessary¹³⁷. This case holds significant importance in comprehending the function of DSPs as overseers of online discourse.

¹³⁴Frosio G and Geiger C (n 9) 31-77.

¹³⁵ *Erbakan v Turkey*, App. no. 59405/00, (ECtHR, 6 July 2006).

¹³⁶ *Garaudy v France* (dec.), App. no. 65831/01, (ECtHR, 24 June 2003). The Court elucidated that Article 17 of the ECHR should be invoked in instances where the contested discourse aims to subvert the true intent of Article 10 of the ECHR by utilising the freedom of expression to advance objectives blatantly at odds with the principles upheld by the Convention. See also *Perinçek v Switzerland App no 27510/08* (ECtHR, 15 October 2015).

¹³⁷ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi'), para. 156.

Delfi, an online news platform, implemented a mechanism for managing comments, comprising several measures, notably an automated filtering system. The automated filtering system comprised the following components: (i) disclaimers concerning accountability for comments and the prohibition of comments contradicting established norms, (ii) an automated deletion system targeting comments containing specific vulgar language, (iii) a notification and removal procedure, and (iv) a practice where administrators proactively remove inappropriate comments at their discretion¹³⁸.

Contents of such components should be promptly eliminated with appropriate and effective measures. This stringent stance was affirmed in the case of *Delfi AS v. Estonia*, which dealt with the belated removal of online comments containing hate speech and calls to violence. The Court emphasized the responsibility of platforms to monitor and promptly address harmful content, including hate speech, to protect individuals. It highlighted that any limitations on freedom of expression must be proportionate and necessary¹³⁹. For example, the title of the news article was “SLK Destroyed Planned Ice Road” which attracted 20 negative comments that promoted hate speech towards SLK Company. This legal precedent holds undeniable significance in grasping the responsibilities of DSPs as guardians of online discourse. The approach adopted by Delfi, an online news platform, entailed implementing a comprehensive mechanism for comment management, encompassing four key measures, among which was the deployment of an automated filtering tool. Article 10 of the ECHR protects freedom of expression. This includes the right to share information and ideas, even if they

¹³⁸*Ibid.*

¹³⁹ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter ‘Delfi’), para. 156. (“Delfi AS: An Estonian news website that allowed readers to comment on articles. Ferry company owner was the plaintiff who sued Delfi over defamatory comments posted by readers on an article about their company. The Issue: Delfi removed the offensive comments after notification but argued that holding them liable for anonymous user comments infringed on their right to freedom of expression. They claimed they shouldn't be considered publishers of the comments. The Court's Decision: The European Court of Human Rights (ECtHR) sided with Estonia by a 15-2 majority. They determined that: Delfi acted as a publisher: Due to their ‘professionally run’ and ‘commercial’ nature, the ECtHR considered Delfi to have editorial control over the comments section, like a publisher. Delfi's measures were inadequate: The court found Delfi's system of relying solely on user reports for comment removal wasn't enough. They expected a more proactive approach like content moderation tools and faster takedowns.”

are offensive. Taking down content can limit free speech. However, the aim is to maintain a balance between freedom of expression and preventing hate speech. Shifting focus to the EU, the regulatory landscape also reflects a commitment to balancing these competing interests. Specifically, the EU aims to remove harmful content while simultaneously protecting people's right to express themselves. Therefore, clearer guidelines are needed to ensure content is removed only when truly necessary and does not stifle legitimate expression. The EU's approach emphasizes the importance of establishing robust, transparent procedures to navigate the complex intersection of free speech and content moderation. Returning to the role of the ECtHR, various justifications for imposing liability on online intermediaries have been delineated by the Court. For example, if platforms have knowledge of illegal content, they have the responsibility of taking it down. This highlights the Court's approach to ensuring that online intermediaries act promptly upon becoming aware of unlawful content, balancing the protection of individual rights with the maintenance of freedom of expression. Although platforms are not required to monitor all content constantly, they must take action to remove illegal content as soon as they discover it or are notified about it. Initially, in the *Delfi* decision, the Court stated that most of the comments constituted 'hate speech or incitement to violence,' thereby deeming them 'patently unlawful' and consequently not warranting the safeguards afforded by Article 10 of the ECHR¹⁴⁰. Secondly, the ECtHR determined that the portal's character as 'professionally run' and 'commercial,' akin to that of a 'publisher,' rendered the measures undertaken by the applicant to eradicate the comments inadequate¹⁴¹. Specifically, the applicant lacked proactive measures to identify and promptly delete harmful content. For instance, the use of algorithms and content moderation policies could have helped identify harmful content more effectively. Furthermore, the Court indicated that while a notice-and-take-down system can be effective, in cases of clearly unlawful comments, such as hate speech and direct threats, liability could still be imposed on the platform even if the comments were removed without delay, particularly if the removal was reactive

¹⁴⁰ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015), para.159

¹⁴¹ *Ibid.*

rather than proactive¹⁴². Finally, the ECtHR endeavoured to circumscribe the broader repercussions of its ruling by specifying its inapplicability to ‘other Internet forums’, such as social media platforms or personal blogs¹⁴³. However, this ruling has prompted significant scholarly critiques suggesting that certain court decisions might not fully align with the principles of freedom of expression. Critics argue that regulating and holding internet intermediaries liable may lead to excessive restrictions on free speech. McGonagle highlights the risk of over-censorship and a chilling effect due to excessive liability on platforms¹⁴⁴. Livingstone notes the tension between protecting against harmful content and maintaining free speech, warning against overly restrictive content moderation¹⁴⁵. Collectively, these scholars emphasise the importance of balancing regulation with the protection of free speech to avoid stifling legitimate discourse¹⁴⁶. The Court's ambiguity in delineating the concept of 'hate speech' has engendered uncertainty among both Internet users and intermediaries¹⁴⁷. This ambiguity arises in two ways: first, the ECtHR does not provide a clear, uniform definition of hate speech; second, the Court considers the context and specific circumstances of each case when determining if speech constitutes 'hate speech'. Furthermore, the lack of precision in defining ‘hate speech’ has prompted intermediaries to devise internal criteria for identifying hate speech, heightening the risk of censorship and curtailing online freedom of expression. Internet intermediaries endeavour to pre-emptively remove potentially defamatory remarks to shield themselves from liability¹⁴⁸. Imposing liability on intermediaries for hate speech could result in platform closures or proactive content filtering, both of

¹⁴² *Ibid.*, para.159 *Delfi AS v Estonia App. no. 64569/09* (ECtHR, 16 June 2015), para.159

¹⁴³ *Ibid.*, para.

¹⁴⁴ McGonagle, T. (2019). The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing. *Human Rights in the Age of Platforms*, pp. 227-254.

¹⁴⁵ Livingstone, S. The *Delfi AS v Estonia* judgement explained. *London School of Economics and Political Science Blog*. (2015) <http://eprints.lse.ac.uk/id/eprint/80766>

¹⁴⁶ McGonagle, 2019, as cited in footnote 145 p.43; Livingstone, 2015, as cited in footnote 146 p.43.

¹⁴⁷ *Delfi AS v Estonia App. no. 64569/09* (ECtHR, 16 June 2015) (hereafter ‘Delfi’), the joint dissenting opinion of Judges Sajó and Tsotsoria, articulated in paragraphs 12 and 13 (pp.70-71), challenges the significance attributed to *Delfi AS*'s commercial and professional characteristics in the context of removing potentially unlawful comments. Moreover, it underscores the potential of comments sections to foster contributions to public discourse independently of centralised media determinations, as delineated in paragraphs 28 and 39, p.77

¹⁴⁸ McGonagle T (n 145) 235

which carry adverse implications for free speech and public discourse online¹⁴⁹. The primary challenge in the ECtHR's jurisprudence concerning online matters lies in a broader uncertainty that extends beyond the CoE's legal scope. This issue mainly centers on the inadequacy in clearly defining 'hate speech'. 'Hate speech' lacks a universally accepted definition and includes a range of hostile messages, from offensive, derogatory, and abusive remarks to intimidating and inflammatory speech that incites violence against specific individuals or groups. Generally, only the most severe forms of hate speech, such as those that incite discrimination, hostility, and violence, are considered unlawful¹⁵⁰. *Delfi AS* stands as the seminal case concerning online hate speech, shedding light on the challenges faced by the ECtHR in addressing this issue. Initially, the Court critiques the insufficient measures taken by the applicant for promptly removing hate speech comments, including the shortcomings in their content monitoring system¹⁵¹. However, this critique could also be construed as an indirect acknowledgment of the complexities associated with defining hate speech in an online milieu, given the multifarious factors involved¹⁵². For example, different contexts of their use and global versus local standards. The ECtHR itself struggles with this definitional constraint by abstaining from delving into the specifics of what rendered certain comments as 'hate speech'¹⁵³. Instead, it simply asserts that 'most' of the disputed comments were categorised as such, without delineating which comments the Court deemed unlawful and the rationale behind such categorisation¹⁵⁴. Shifting focus to the European Union system, the CJEU's approach to defining hate speech requires a clearer intention to target specific groups. This doctrinal divergence highlights the complexity and inconsistency in legal interpretations between the ECtHR and the CJEU regarding

¹⁴⁹McGonagle T (n 145) 236. See also *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi').

¹⁵⁰ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi'); See Report of the Special Rapporteur on minority issues, Rita Izsák (A/HRC/28/64), Human Rights Council, Twenty-eighth session.

¹⁵¹ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi').

¹⁵²*Ibid.*

¹⁵³*Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi').

¹⁵⁴*Ibid.*

what constitutes hate speech. Based on a doctrinal interpretation of the CJEU case law, McGonagle argued that statements must exhibit an intention to target a 'member of a specific group' to qualify as hate speech¹⁵⁵. However, in *Delfi AS* hate speech arises from the absence of reference to a particular group, instead using vague adjectives such as 'hateful' associated with 'unlawful comments' and 'incitement to violence'¹⁵⁶. The fluctuating and 'indeterminate' depiction of hate speech in *Delfi AS* persisted in the subsequent MTE decision¹⁵⁷. Although the MTE decision was successful in limiting the extent to which intermediaries (like online platforms) are held liable for user-generated content, it continued to display an inconsistent and unclear portrayal of what constitutes hate speech¹⁵⁸. In the *MTE v. Hungary* case, the ECtHR engaged with the aftermath of the *Delfi AS* case to elucidate its implications¹⁵⁹. MTE published an opinion piece critical of real estate company's business practices on their website. Index.hu, a news portal, also allowed comments on an article related to the company. Some user comments on both platforms contained defamatory remarks about the real estate company. Hungarian courts ruled that MTE and Index.hu were liable for the comments, despite not writing them themselves. Within Hungarian legal proceedings, the two websites faced strict liability for remarks made by a third-party user. These remarks were categorised as 'value judgments' regarding a 'dispute over the commercial policy of the plaintiff company', notwithstanding their 'vulgar' and 'offensive' nature¹⁶⁰. MTE underscored that the liability of internet intermediaries hinges solely on instances of hate speech and individual threats to physical safety¹⁶¹. The ECtHR differentiated the *Delfi AS* case from the present scenario by accentuating the public interest in MTE's comments and

¹⁵⁵ McGonagle (n 145) 245.

¹⁵⁶ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi')

¹⁵⁷ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016)

¹⁵⁸ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter 'Delfi')

¹⁵⁹ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016), para 91.

¹⁶⁰ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016), para.75

¹⁶¹ *Ibid.*, para. 91

their lack of 'clearly illegal' status¹⁶². Additionally, the Court criticised the inflexibility of Hungarian laws, which hindered the balancing of freedom of expression and privacy rights. For instance, inflexible laws didn't allow online platforms like MTE and Index.hu to take down comments in a nuanced way. They were forced to choose between keeping all comments up, potentially violating someone's reputation, or taking everything down, which could stifle free speech. Another important consideration is the context. The context in which a comment is made significantly influences the assessment of whether it is harmful or acceptable when moderating content. For example, if the comments were simply critical opinions rather than malicious attacks, the laws might not have allowed for any distinction. A more internet-friendly stance was evident in the *Yıldırım* case¹⁶³. The joint dissenting opinion of Judges Sajó and Tsotsoria, articulated in paragraphs 12 and 13, challenges the significance attributed to Delfi AS's commercial and professional characteristics in the context of removing potentially unlawful comments. Moreover, it underscores the potential of comments sections to foster contributions to public discourse independently of centralised media determinations. The Court acknowledged internet intermediaries as participants in public discourse¹⁶⁴ and rejected for them to implement proactive filters against potentially unlawful comments¹⁶⁵. This decisive shift was accompanied by five pertinent criteria for evaluating the proportionality of interference in platform liability cases¹⁶⁶. These criteria include the context and content of disputed comments, the liability of comment authors, the measures taken by website operators, the behaviour of the injured

¹⁶² *Ibid.*, paras. 64-70

¹⁶³ *Ahmet Yıldırım v Turkey* App No 3111/10, (ECtHR, 18th December 2012).

¹⁶⁴ *Ibid.*, para.54

¹⁶⁵ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016), para 82.

¹⁶⁶ These criteria typically apply to cases involving platform liability and proportionality assessments but are not applicable to instances of hate speech or encouragement of violence due to the distinct legal standards and heightened scrutiny such content warrants under both national and international law. The context of platform liability and the evaluation of proportionality are complex and often vary depending on the nature of the content in question. For example, while general content may be evaluated based on proportionality criteria such as necessity, suitability, and minimal intrusion, content involving hate speech or violence typically requires immediate and stringent measures due to its potential harm and legal implications.

party, and consequences for both the injured party and the complainants¹⁶⁷. The established criteria have provided clarity to the continually evolving legal landscape, and other cases have been adjudicated using similar criteria¹⁶⁸. Furthermore, the ECtHR in the MTE case asserted that the applicants' privacy rights had not been infringed, as the comments were merely defamatory and did not meet the threshold for hate speech or incitement to violence. This rationale revealed the Court's capacity to adapt and reconsider its approach to online freedom of expression in response to academic and legal critiques. The distinction in speech between the *Delfi* and MTE cases is attributed to the characterisation of MTE's remarks as 'offensive and vulgar', whereas those in the *Delfi AS* case were deemed to constitute hate speech or incitement to violence¹⁶⁹. Although there exists definition of the speech, the *Yıldırım* case provided further indications related to the important principle of proportionality that can justify a defence. Overall, the ECtHR's jurisprudence underscores the complex interplay between freedom of expression and the need to address hate speech online, requiring a delicate balance to uphold both principles effectively. The *Delfi AS*, *MTE v. Hungary*, and other landmark cases highlight the challenges and evolving standards in defining and regulating hate speech, demonstrating the Court's efforts to protect fundamental rights while addressing harmful content. Furthermore, the Court of Justice of the European Union (CJEU) has significantly influenced this area, particularly through its rulings in cases such as *Google France* and *L'Oréal v. eBay*. The CJEU's approach to intermediary liability emphasises the need for a clearer intention to target specific groups for speech to be considered hate speech, and it has developed a framework that balances regulation with the protection of free speech. This evolving legal landscape highlights the ongoing dialogue between these courts and their efforts to navigate the complexities of digital communication

¹⁶⁷ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016), paras 72-86.

¹⁶⁸ See, for instance: *Pihl v Sweden* App No 74742/14 (ECtHR, 9th March 2017); *Tamiz v the United Kingdom* App No 3877/14 (ECtHR, 9th September 2017); *Høiness v Norway* App No 43624/14 (ECtHR, 9th March 2019).

¹⁶⁹ *Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary* App No 22947/13 (ECtHR, 2nd February 2016), para. 64

and content moderation. The next section will delve into the evolving role of the CJEU in shaping intermediary liability and freedom of expression online, examining key cases and their implications for the regulatory framework.

2.7 The Evolving Role of the CJEU in Shaping Intermediary Liability and Freedom of Expression Online

The CJEU has wielded substantial influence in shaping the practical implementation of the liability exemption framework for intermediaries. Initially, the focus on the liability of Internet intermediaries revolved around ensuring a 'fair balance' between the internal market, fundamental rights such as the right to freedom of expression, and trademark rights. This balance is about protecting the economic interests and smooth functioning of the internal market while also safeguarding fundamental rights and addressing the intellectual property concerns of trademark holders¹⁷⁰. However, pivotal cases such as *Google France* and *L'Oréal v. eBay*¹⁷¹ have refined the concept of intermediary liability by introducing a distinction between 'active' and 'passive' service providers. In the *Google France* case, the CJEU acknowledged that Google did not play an 'active' role allowing it to 'know or control the data stored,' thereby exempting it from liability¹⁷². Subsequently, in the *L'Oréal* case, the Court elaborated on this distinction, ruling that even 'passive hosts' could be subject to injunctions, notwithstanding their lack of obligation to actively monitor all data, which would contravene applicable law¹⁷³. The *L'Oréal* case holds relevance for several reasons. Firstly, the CJEU emphasised the need for any injunctions against intermediaries to be proportionate and specific, ensuring that they do not impose a general monitoring obligation or disproportionately interfere with fundamental rights, such as freedom of expression. This stance is akin to the one adopted by the European Court of Human Rights in *Ahmet Yildirim v Turkey* App No 3111/10, where a broad and non-specific blocking order

¹⁷⁰ C-275/06 *Promusicae v Telefónica de España SAU* [2008] ECR I-271 para 70; C-557/07 *LSG v Tele2* [2009] ECR I- 1227 para 28.

¹⁷¹ Joined Cases C-236/08 to C-238/08 *Google France and Google Inc. v Louis Vuitton Malletier SA* and others [2010] ECLI:EU:C:2010:159

¹⁷² C-236/08 *Google France v Louis Vuitton Malletier SA* [2008] ECLI:EU:C:2010:159 para 120.

¹⁷³ C-324/09 *L'Oréal SA v eBay International AG* [2011] ECR I-474, para.127.

was deemed a disproportionate interference with freedom of expression¹⁷⁴. Secondly, the Court acknowledged the role of platforms like eBay in facilitating user-uploaded advertisements and recognized the increasing significance of online commerce. The Court highlighted the importance of ensuring that such platforms take appropriate measures to prevent the sale of counterfeit goods while balancing the interests of various stakeholders¹⁷⁵. In the case *L'Oréal v. eBay* (C-324/09), the plaintiff, L'Oréal, a cosmetic company accused the defendant, eBay, an online marketplace, of being liable for the sale of counterfeit and unauthorised L'Oréal products on its platform. The CJEU ruled that eBay could be held accountable based on two main circumstances including knowledge of infringement and an active role. Specifically, the Court held that eBay's actions of using trademarks in advertising as going beyond a passive role and contributing to the infringement. They claimed eBay infringed their trademarks by displaying their trademarks on the website using keywords containing their trademarks for advertising on search engines in relation to these infringing goods. Additionally, L'Oréal argued that eBay was jointly liable with the individual sellers for the infringements. In the case *Google France v. Louis Vuitton Malletier* (Joined Cases C-236/08 to C-238/08), Louis Vuitton accused Google of trademark infringement through its AdWords program. Louis Vuitton argued that Google was infringing on their trademarks by allowing competitors to use their trademarks as keywords to trigger ads and displaying these ads alongside search results for genuine Louis Vuitton products. The CJEU ruled in favour of Google, stating that search engines themselves don't infringe trademarks by allowing advertisers to use competitor trademarks as keywords. Google's role was considered neutral and technical, as it simply displayed ads based on user searches and advertiser bids. This decision reflects the CJEU's effort to balance the protection of trademark rights with the need to maintain a fair and functional digital marketplace, ensuring that intermediary liability does not unjustifiably impede on the freedom of commercial expression and innovation. The contrasting strategies of the CJEU, which aims to cultivate a robust online marketplace, and the ECtHR, which

¹⁷⁴ *Ahmet Yildırım v Turkey* App No 3111/10, (ECtHR, 18th December 2012).

¹⁷⁵ C-324/09 *L'Oréal SA v eBay International AG* [2011] ECR I-474, paras 155-159.

is dedicated to protecting fundamental human rights, especially freedom of expression, highlight the differing methods each institution uses to address the responsibilities of online intermediaries while ensuring the protection of fundamental rights. In the case of *Sotiris Papasavvas v O Fileleftheros Dimosia Etairia Ltd* (C-291/13), the liability exemption wasn't directly applied because the lawsuit wasn't against an intermediary service provider. The E-commerce Directive grants exemptions from liability for certain information society service providers, like social media platforms, under specific conditions (explained in the previous section). *O Fileleftheros Dimosia Etairia Ltd*, the publisher, wasn't considered an intermediary under the E-commerce Directive. The E-commerce Directive focuses on online service providers, providing liability exemptions under certain conditions, while *O Fileleftheros*, a publisher, was producing content both in a physical newspaper and online. This distinction is important because the Directive's liability exemptions apply specifically to online intermediaries, not to traditional publishers who exercise editorial control over their content. The CJEU ruled that newspaper publishers don't qualify as 'mere conduit' or 'caching' providers because they have editorial control over the information they publish. This showed that the Court considered the role played by the service provider cementing the importance of the active/passive distinction as seen in case *Google France* case¹⁷⁶. This decision made in the *Google France* case is noteworthy for two fundamental reasons. First, it ensures that information society services, including those operating on a 'freemium' model, are considered remunerated services even if they are not directly paid for by the recipients. Freemium services offer a basic version for free but have premium features that users can pay for. Many online services like social media platforms, games, and cloud storage follow this model. The term 'remunerated' refers to services where users pay a fee for access or use, or where the service is funded through advertising revenue. This inclusion is crucial given the growth of online business models that rely on donations and advertising revenue for financial gain¹⁷⁷. Moreover, the

¹⁷⁶ C-236/08 *Google France v Louis Vuitton Malletier SA* [2008] ECLI:EU:C:2010:159 para 120.

¹⁷⁷ Schreiner, M., & Hess, T. (2013). On The Willingness To Pay For Privacy As A Freemium Model: First Empirical Evidence. *ECIS 2013 Research in Progress*, pp.

frequency with which users upload information on these services means that they should *prima facie* fall within the scope of the E-Commerce directive¹⁷⁸. Secondly, this is an interesting case to compare with the landmark *Delfi AS* ruling of the ECtHR, which addressed the liability of an online news portal for user-generated comments and highlighted different approaches to intermediary liability, as we have analysed in the specific Delfi case in the previous paragraph¹⁷⁹. The contrasting rulings show how courts weigh various factors to determine when platforms should be held responsible and how to maintain a vibrant online space with some level of accountability. These cases illustrate that while the CJEU has emphasized the importance of freedom of expression in the context of online communications, there is a growing recognition of the need for accountability and regulation to prevent abuse. The Delfi case could be seen as an indication of a shift towards more stringent measures to hold platforms accountable, highlighting a potential reevaluation of the balance between freedom of expression and the responsibility of online intermediaries¹⁸⁰. It is noteworthy that the E-Commerce Directive and the liability exemptions stipulated in Articles 12-15 were subject to comprehensive examination in the cases of C-314/12 *UPC Telekabel Wien* and C-484/14 *McFadden*, illustrating the ongoing judicial scrutiny of intermediary liability.

In these cases, the Court further scrutinised the delicate balance between copyright enforcement, trademark rights, and the freedom of internet access providers to conduct their business while also considering fundamental rights such as freedom of expression. These cases illustrate how the Court navigates the intersection of these complex and often competing interests. However, as this matter does not directly pertain to the primary focus on freedom of expression, it will not be further elaborated upon. The Court has established important principles that both protect freedom of expression and encourage online platforms to take some responsibility for the content they host. After

¹⁷⁸ This hinges upon whether they are categorised as passive or active entities in accordance with the directives outlined in the E-Commerce Directive and the subsequent case law elucidated by the Court of Justice of the European Union (CJEU).

¹⁷⁹ *Delfi AS v Estonia* App No 64569/09 (ECtHR, 16th June 2015)

¹⁸⁰ C-314/12 *UPC Telekabel Wien* [2014] ECLI:EU:C:2014:192 and (2) C-484/14 *McFadden* [2014] ECLI:EU:C:2016:689.

assessing this pro-freedom-of-expression jurisprudence of the CJEU, we can see that the EU's first steps in addressing this difficult balancing act have been valuable. However, the 'Articles 12-15' approach of the E-commerce Directive dealing with intermediary liability and safe harbour, has become increasingly unpopular with the EU Commission, the CJEU and even some national legislatures, especially considering the growing importance of ISPs and social media. In line with the judicial scrutiny of intermediary liability under the E-Commerce Directive observed in the *Papasavvas* case, the *Eva Glawischnig-Piesczek v Facebook* case also addresses the issue of defamatory comments posted on an Internet service provider's platform, specifically Facebook¹⁸¹. In the *Eva Glawischnig-Piesczek v Facebook* case, the comments were attached to a journalistic article that an anonymous user had shared publicly on Facebook. The defamatory comments concerned a high-profile Austrian politician¹⁸². The questions submitted to the CJEU focused on the scope of the injunction, considering the general prohibition of monitoring of E-Commerce Directive¹⁸³. First, the decision raises concerns about forum shopping, allowing plaintiffs to bring defamation actions in plaintiff-friendly jurisdictions, knowing that they may result in global removals¹⁸⁴. Second, the CJEU's failure to define 'equivalent' information means that it can have different meanings in different courts. While it is commendable that AG Szpunar¹⁸⁵ acknowledged the 'interpretative difficulties' inherent in not defining 'equivalent information', it is nevertheless regrettable that the final decision refused to fully address this issue. It could also be argued that the CJEU's hand was forced because

¹⁸¹ C-18/18 *Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821

¹⁸² Opinion of AG Szpunar at para 12: "...That user also published, in connection with that article, an accompanying disparaging comment about the applicant accusing her of being a 'lousy traitor of the people', a 'corrupt oaf' and a member of a 'fascist party'. The content placed online by that user could be consulted by any user of the platform in question".

¹⁸³ Council and Parliament Directive 2000/31/EC of 8th June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (E-Commerce Directive) [2000] OJ L 178/1, Article 15(1).

¹⁸⁴ C-18/18 *Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821

¹⁸⁵ AG Szpunar refers to Advocate General Maciej Szpunar, a legal advisor to the Court of Justice of the European Union (CJEU). The role of an Advocate General is to provide independent and impartial opinions on cases before the Court. These opinions are not binding but are influential and often considered by the judges of the CJEU when making their decisions.

the separate liability regime within the E-Commerce Directive was necessary to address a traditionally national competence. Apart from the defamation issues, the decision could also be criticised for placing 'false confidence' in the usefulness of automatic detection systems; these systems are not yet up to the task of detecting 'equivalent' information in nuanced contexts across all social media posts¹⁸⁶. Thus, it is argued that this case potentially increases the responsibilities of platforms to proactively monitor and remove defamatory content, which could lead to broader content removal practices to evade potential liability for defamatory content. Notwithstanding these criticisms, it should be noted that the CJEU decided this case in the absence of a clear EU legal framework for defamation¹⁸⁷. These cases signify a shift in the CJEU's consideration of freedom of expression in the context of online communications, emphasising the delicate balance between copyright enforcement and the freedom of internet access providers to conduct business. However, unresolved issues remain regarding the definition of 'equivalent information' and the efficacy of automatic detection systems, as well as the need for a more nuanced balancing of fundamental rights by the CJEU. Overall, the ECtHR's jurisprudence underscores the complex interplay between freedom of expression and the need to address hate speech online, requiring a delicate balance to uphold both principles effectively. The *Delfi AS, MTE v. Hungary*, among other landmark rulings, highlight the challenges and evolving standards in defining and regulating hate speech, demonstrating the Court's efforts to protect fundamental rights while addressing harmful content. Furthermore, the CJEU has also significantly influenced this area, particularly through its rulings in cases such as *Google France* and *L'Oréal v. eBay*. The CJEU's approach to intermediary liability emphasises the need for a clearer intent to target specific groups for speech to be considered hate speech, and it has developed a framework that balances regulation with the protection of free speech. This evolving legal landscape highlights the ongoing dialogue between these Courts and their efforts to navigate the complexities of digital

¹⁸⁶ C-18/18 *Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821

¹⁸⁷ Pollicino, O., & Bassini, M. (2014). Free speech, defamation and the limits to freedom of expression in the EU: a comparative analysis. In *Research Handbook on EU Internet Law* (pp. 508-542). Edward Elgar Publishing

communication and content moderation. The next section will delve into the evolving role of the CJEU in shaping intermediary liability and freedom of expression online, examining key cases and their implications for the regulatory framework.

2.8 The European Court of Human Rights and the Council of Europe on the responsibility of intermediaries and the protection of human rights in the digital world

In carrying out this analysis, it is essential to scrutinise the disparity between the EU E-Commerce Directive of 2000 and the stance adopted by the ECtHR in the *Delfi AS* case¹⁸⁸. This scrutiny is crucial because it highlights the differences in how the EU and the ECtHR approach the liability of internet intermediaries. Understanding these differences is fundamental to answering the research question: How do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression? By examining these contrasting approaches, we can better understand the challenges and complexities involved in developing a coherent and effective regulatory framework that protects freedom of expression while addressing the harms of online hate speech. The majority position of the ECtHR judges dismissed the applicability of the passive intermediary exemption outlined in Article 14 of the Directive to *Delfi AS*¹⁸⁹. This dismissal was based on the assertion that the company actively engaged in a system characterised by 'notice-and-takedown' procedures. This comparison highlights two notable aspects. Firstly, there is a debate regarding how to categorise internet companies, with some arguing that categorisation should not solely be based on their actions (active vs. passive). This notion is largely supported by the judges in the *Delfi AS* case, where they highlighted the complexities of determining liability based solely on such distinctions¹⁹⁰. Secondly, the majority stance in *Delfi AS* appeared to inadequately grasp the significance of 'actual knowledge' – meaning that an intermediary

¹⁸⁸ P. Przemysław Polański, Rethinking the notion of hosting in the aftermath of *Delfi*: Shifting from liability to responsibility?, *Computer Law & Security Review*, Volume 34, Issue 4, 2018, Pages 870-880, ISSN 0267-3649, <https://doi.org/10.1016/j.clsr.2018.05.034>.

¹⁸⁹ P. Przemysław Polański (n 189) 870-880

¹⁹⁰ This perspective is supported by the 2018(2) Council of Ministers Recommendation as well as by the European Commission in its endeavours to embrace a novel framework for intermediary liability.

service provider like *Delfi AS* has clear and specific information that illegal content is present on their platform- within the framework of the 2000 E-Commerce Directive. This suggests that a requirement to eliminate illicit content in the absence of genuine awareness of its presence could potentially impose strict liability on intermediaries¹⁹¹. The 2000 E-Commerce Directive established a legal framework for online services within the European Union, aiming to reduce barriers for e-commerce, protect consumers, and define intermediary liability. One crucial condition for intermediary liability exemption was not having actual knowledge of illegal content¹⁹². The divergence in approach between the CJEU (fostering a thriving online marketplace) and the ECtHR (protecting fundamental human rights, particularly freedom of expression) underscores variations in addressing the obligations of online intermediaries while safeguarding fundamental human rights. Conversely, the EU primarily focuses on expeditiously removing illicit content without providing explicit guidance on reconciling this goal with adherence to Article 10 of the ECHR¹⁹³. This ambiguity poses potential negative consequences for both internet users and intermediaries. For users, the lack of a clear-cut definition complicates determining whether a directed comment qualifies as hate speech, thus falling outside the protective scope of Article 10 of the European Convention on Human Rights. While, the ECtHR provides a framework, intermediaries require more specific guidance to avoid overstepping their bounds when removing content. Consequently, this responsibility may exert undue pressure on intermediaries and their moderators, necessitating the interpretation of the demarcation between permissible and impermissible speech without sufficient legal expertise to reconcile conflicting fundamental human rights. Uncertainty regarding intermediary liability for hate speech may lead platforms to cease operations to mitigate liability risks or resort to blanket content filtration, thereby curtailing forums for public discourse. Ultimately, the influence of the ECHR is limited by its indirect application to private entities, despite the significant influence wielded by major technology

¹⁹¹ *Delfi AS v Estonia* App. no. 64569/09 (ECtHR, 16 June 2015) (hereafter ‘Delfi’).

¹⁹² Höning, Nina. "The European Directive on e-Commerce (2000/31/EC) and its Consequences on the Conflict of Laws." *Global Jurist Topics* 5.2 (2005).

¹⁹³ ‘EU Code of conduct on countering illegal hate speech online’ (EU Commission, 30th June 2016).

corporations, which often surpasses that of national governments. These corporations largely formulate their own regulations, albeit under the tacit oversight of the EU¹⁹⁴. This raises questions about the justification of case law. However, ECtHR case law still significantly impacts three aspects: (a) rulings on cases involving government restrictions on free speech can inform how courts and governments approach similar situations with private companies; (b) the ECtHR's reasoning and principles can shape 'soft law', influencing how companies shape their policies to avoid potential future legal challenges or public criticism for violating human rights principles; (c) high-profile ECtHR rulings can raise public awareness and pressure companies to act in line with human rights standards. While the positive obligation of states to uphold the human rights of individuals may be activated when the actions or inactions of a private intermediary lead to a human rights violation, the Convention would undoubtedly have greater efficacy if major platforms were directly held accountable¹⁹⁵.

2.9 Conclusion

The evolution of the Internet has created a complex landscape where the potential harm of hate speech conflicts with the principles of freedom of expression, challenging legal frameworks to find a delicate balance between these competing interests. This chapter elucidates two primary doctrines used to address online speech: non-exceptionalism, which applies existing laws from the physical world to the online realm, and exceptionalism, which involves crafting new regulations tailored to the unique characteristics of online communication. However, it is imperative to strike a balance between these approaches, as excessive regulation risks censoring legitimate speech, while insufficient regulation allows hate speech to proliferate. Analysing legislative responses to online hate speech demands a meticulous examination within the intricate web of regulations governing platform liability and fundamental rights. Over time, attempts by national governments, supranational organisations, and international bodies to establish regulations concerning intermediary and platform liability have

¹⁹⁴ 'EU Code of conduct on countering illegal hate speech online' (EU Commission, 30th June 2016)

¹⁹⁵ McGonagle (n 145) 246.

grappled with reconciling divergent rights impacted by the activities and obligations of digital service providers. In examining the nuanced balance between combating online hate speech and upholding the fundamental right to freedom of expression, this analysis highlighted the evolution of regulatory frameworks within the European Union, emphasising the pivotal role of intermediary service providers in managing unlawful content. The analysis underscored the EU's approach, particularly through the E-Commerce Directive and the DSA, which aims to navigate the tension between ensuring user safety and preserving free speech online. Key judicial rulings from the ECtHR and the CJEU have further shaped this landscape, emphasising the need for intermediaries to act upon direct knowledge of illicit content while maintaining the delicate balance of protecting fundamental rights. As the digital world continues to evolve, the ongoing challenge will be to refine these frameworks to effectively address online hate speech without unduly compromising freedom of expression. Recent European legislative measures, such as the notice-and-takedown system (NTDS), introduce a multi-tiered structure of prerequisites and protocols. The Notice-and-Takedown System is a legal framework used to address online content that allegedly infringes on someone's rights, such as copyright violations, defamation, or other illegal content. This system includes the complainant's genuine belief in rights infringement, clear identification of infringing content and its location, and provision of legal justification for takedown. The process involves submitting a notice to the intermediary, who must quickly process it and remove the content if the criteria are met. Additionally, there are protocols for counter-notices to dispute claims. Intermediaries are required to have direct knowledge of illicit content and promptly remove it within specified time limits. Legal immunity for intermediaries under the “safe harbour model” is contingent upon meeting these specific requirements rather than being automatically granted. The notice-and-action system, which provides a more comprehensive approach, extends the notice-and-takedown concept by adding more steps and responsibilities for both the complainant and the intermediary. This system often includes enhanced procedures, requiring intermediaries to not only remove content but also notify the content creator and provide mechanisms for dispute resolution. Furthermore, it mandates transparency reports and

robust mechanisms for resolving disputes between complainants and content creators. Intermediaries must comply with specific regulatory standards to maintain their safe harbour protections, including implementing effective notice-and-action procedures and ensuring transparency. The Notice-and-Takedown System (NTDS) is primarily implemented through the E-Commerce Directive, while the more comprehensive Notice-and-Action System is part of the Digital Services Act (DSA). Consequently, the EU adopts the “safe harbour model,” wherein legal immunity is contingent upon meeting specific requirements rather than being automatically granted. The jurisprudence of the CJEU and the ECtHR has established that intermediary liability is contingent upon the knowledge and active role of the intermediary in relation to the illicit content. Both courts play critical roles in addressing the balance between protecting fundamental rights, such as freedom of expression, and ensuring effective measures against illegal content. The ECtHR focuses on safeguarding human rights, often emphasising the protection of free expression while setting clear boundaries against hate speech. Conversely, the CJEU primarily aims to ensure the proper functioning of the internal market, adopting a more structured approach to intermediary liability that requires intermediaries to act upon direct knowledge of illegal content. These differing perspectives highlight the complex interplay between freedom of expression and regulatory measures in the digital age.

Chapter 3: Assessing the “Systemic Risks” in Digital Services Act and International Human Rights Law

Summary: 3.1 Introduction – **Part 1 – EU Law:** 3.2 The provisions of Digital Services Act in response to hate speech – 3.2.1 Transparency Measures – 3.2.2 Systemic Risks in the DSA – 3.2.3 Notice and Action Process in the DSA – 3.2.4 Conditional Liability Framework – 3.2.5 Automated Content Moderation and its Risks – 3.2.6 The Digital Services Coordinators in the DSA – 3.2.7 Conclusion - 3.3 The First Systemic Risk: Dissemination of Illega Content - Defining and Adressing ‘Illegal Content’ in the DSA - 3.4 The Second Systemic Risk: The Actual or Foreseeable Negative

Effects on the Exercise of Fundamental Rights - 3.5. Third and Fourth Categories of Systemic Risks: Manipulation of Very Large Online Platforms (VLOPs) and Search Engines - Effects and Consequences - 3.6 Mitigation of systemic risks

–**Part 2: International Law**- 3.7 International human rights law (IHRL) as a Framework for the Moderation of Online Content – 3.8 Hate speech and International Human Rights Law - 3.9. Conclusions

3.1 Introduction

This chapter aims to reflect on intricate legal matters pertaining to the responsibilities associated with addressing online hate speech. It explores these issues through the lens of EU legal frameworks while acknowledging the evolving landscape of technology. The internet, with its vast user base of 4.74 billion individuals, serves as a significant conduit for the dissemination of hate speech, often perpetuated by influential figures and celebrities like former US President Donald Trump. Such speech can have detrimental effects, fostering resentment and hate towards targeted groups, particularly immigrants. Consequently, it is imperative to understand the legislative framework delineating the boundaries between legal and illegal actions, specifically discerning what constitutes hate speech in online contexts. The enactment of the Digital Services Act (DSA) on 17 February 2024 necessitates careful examination due to the central role of online platforms in contemporary society and the potential harm inflicted by online content¹⁹⁶. In terms of centrality, the DSA protect internet and online platforms, which have become important aspects of our lives. Therefore, the subsequent analysis aims to achieve two primary objectives: firstly, to examine the moderation of online content, and secondly, to evaluate how the Digital Services Act (DSA) navigates the tension between safeguarding freedom of expression online and the escalating demands for content removal, which could undermine free speech and potentially diminish civic space. The DSA proceeds to scale down

¹⁹⁶ 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). *See infra*, para. 2

the powers of platforms through the procedural harmonisation of content control, the establishment of *ad hoc* national and supranational public bodies and the conferral of specific enforcement powers to the Commission. In this way, the propensity to delegate public functions to the private companies owning the platforms is reduced, though not eliminated. Against this backdrop, this chapter will provide an overview of the challenges arising from the current legal framework, as defined by EU law. The analysis will place particular emphasis on the contentious issues related to managing freedom of expression, with a stronger focus on online contexts rather than offline ones. Additionally, it will be crucial to ensure effective protection against the dissemination of hate speech across social networks, news platforms, videos, and other online media¹⁹⁷. In particular, the main aim of this chapter is to explore the current legal framework within the EU to understand whether and how the DSA, balances the responsibility of Internet intermediaries in the moderation process, and the protection of free speech from censorship. Finally, this analysis proposes the adoption of International Human Rights Law (IHRL) as the most effective framework for regulating online content moderation. Although private platforms are not directly bound by international human rights norms, the examination of EU law, the Council of Europe definitions, and the jurisprudence of the European Court of Human Rights (ECtHR) and the Court of Justice of the European Union (CJEU) supports the view that the International Covenant on Civil and Political Rights (ICCPR) offers a comprehensive framework for addressing these issues. However, it is important to note that this thesis does not delve deeply into the application of IHRL in this context (see Sections 3.7 and 3.8 for further

¹⁹⁷The DSA therefore aims to revise the rules governing so-called user-generated content and the related liability of providers with reference to profiles relating to violence, fake news, hate, as well as violations of intellectual property rights. To achieve this, the Digital Services Act (DSA) seeks to establish distinct 'categories' of digital service providers and accordingly adjust their liability based on their awareness of user-uploaded content. The proposal seems to develop, without innovating in a revolutionary way, concepts already covered by EU law: it, for instance, distinguishes mere providers who only provide connectivity services and therefore have little or no capacity to moderate the content disseminated by their customers, from providers such as Facebook or Google who, instead, have far more penetrating knowledge - and therefore moderation - capacities, and aggravates the liability of the latter. On this subject, see 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 (E-Commerce Directive).

discussion). The analysis essentially aims at highlighting the need to regulate the complex phenomenon of fundamental rights in the digital society by balancing the need to protect the rights of the individual and ensuring technological progress. This objective is intricately linked to the main research question: "How do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression"? Each legal framework seeks to navigate the delicate equilibrium between safeguarding individuals from harmful online content and upholding the principles of free speech. By examining these frameworks, the analysis reveals the varied strategies employed to impose liability on internet intermediaries, ensuring that regulatory measures do not stifle innovation while protecting fundamental human rights in the digital age. This comprehensive approach underscores the importance of a balanced regulatory environment that both mitigates online harms and fosters technological advancement. In relation to the central theme of this chapter, which examines the impact of the Digital Services Act (DSA) on fundamental rights, particularly freedom of expression, it is essential to first understand the nature of the so-called 'systemic risks.' While the DSA addresses a broad range of systemic risks, this analysis focuses on those most pertinent to our research question: "How do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression?" These specific risks include: 1) illegal content such as hate speech, 2) content affecting fundamental rights, including those restricting freedom of expression, 3) content related to gender-based violence, and 4) content affecting civic discourse¹⁹⁸. These risks are deemed particularly dangerous due to their potential to cause widespread and significant harm within the EU. Recitals 79 and subsequent sections, along with Article 34(1) of the DSA Regulation, specifically address the new obligations imposed on Very Large Online Platforms (VLOPs) concerning the assessment and mitigation of these systemic

¹⁹⁸ Broughton Micova, S. (2021). *What is the harm in size?* Centre on Regulation in Europe (CERRE), pp.12-17.

risks¹⁹⁹. In the long run, educational effort must be made in addition to legislation on an international basis, to provide a free marketplace for ideas where the collective renders a judgment on the acceptability of ideas and hate speech is naturally suppressed under proper manners when utilising freedom of expression. As a result, ideas including offensive speeches that may carry underlying beneficial intentions can be expressed freely. To ascertain whether decisions grounded in human rights considerations surpass those derived from computational calculations, it is imperative to comprehend human decision-making processes. The application of conceptual frameworks, such as algorithms, has proven to be distinctive. For instance, humans are capable of recognising instances where a rule should not be applied, unlike algorithms. To enhance decision-making, it is crucial to gain a deeper understanding of how algorithms function, as well as how they are designed and implemented²⁰⁰. The Digital Services Act requires very large online platforms (VLOPs) and very large online search engines (VLOSEs) to conduct thorough risk assessments concerning the design of their algorithmic systems. They must also implement measures to mitigate these risks, particularly in terms of how their services impact the exercise of fundamental rights²⁰¹. This includes modifying the design, features, and functionality of their services and online interfaces, adjusting content moderation processes, and allocating appropriate resources for decision-making and content moderation tasks²⁰². A more comprehensive framework is essential for regulating these technologies, clearly outlining the responsibilities and liabilities of service providers²⁰³. The chapter is structured

¹⁹⁹Recital 79 and art. 34 (1), 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).

²⁰⁰ Frosio G and Geiger C (n 9) 33-36.

²⁰¹Art.16 and Recital 81, 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).

²⁰²Art.35, 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).

²⁰³European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts, n.30.

into two distinct sections. The first part discusses the DSA's role in regulating digital activities, focusing on transparency and accountability measures for content moderation. It highlights the requirements for platforms to justify content moderation decisions, establish complaint mechanisms, and inform users about content removal. It also examines the systemic risks defined in the DSA, such as illegal content and infringements on fundamental rights, and the specific obligations for VLOPs to assess and mitigate these risks. Additionally, the chapter explores the use of artificial intelligence (AI) for content moderation and the potential impact on freedom of expression. The second part addresses the effects of the DSA on fundamental rights, assessing risks related to privacy, non-discrimination, and public safety. It discusses the manipulation of VLOPs, and search engines and the challenges posed by algorithmic decision-making. The chapter also explores risk mitigation measures outlined in Article 35 of the DSA, emphasising the importance of reasonable, proportionate, and effective measures. Furthermore, the chapter advocates for adopting IHRL principles for content moderation, arguing that this approach can better balance regulation and freedom of expression. It examines the benefits and challenges of applying IHRL to private platforms and explores the regulation of online hate speech under IHRL, focusing on defining hate speech according to the ICCPR and the Rabat Action Plan. The chapter emphasises the importance of balancing offensive content control with protecting free speech.

Part 1 – EU Law

3.2 The provisions of Digital Services Act in response to hate speech

In this section, we will analyse the obligation to remove unlawful content within the regulatory framework of the Digital Services Act (DSA). The DSA assumes a central role in regulating activities concerning 'information' in the digital realm. This new legislative framework aims to combat the dissemination of 'illegal' content online and enhance the transparency of large technology companies (referred to as big-tech companies), thereby shaping a 'new digital world' with far-reaching societal

implications²⁰⁴. By introducing sector-specific rules²⁰⁵, the DSA seeks to ensure the adequate protection of fundamental rights while imposing transparency and accountability measures on platforms and delineating new rights for users. This section begins by introducing the DSA and its central role in regulating online content, particularly in combating illegal content and enhancing transparency among large technology companies. It highlights the potential global impact of the DSA on content moderation and fundamental rights protection. The section then delves into the transparency measures mandated by the DSA, explaining the obligations placed on platforms to justify content moderation decisions, inform users about content removal, and implement internal complaint mechanisms, emphasising the importance of transparency in protecting freedom of expression. Next, it examines the systemic risks associated with the DSA, focusing on three primary risks: the dissemination of illegal content, impacts on fundamental rights, and manipulation and public safety risks. This section underscores the significance of these risks in the context of hate speech and content moderation. The discussion then moves to the 'notice and action' process established by the DSA, outlining the procedural requirements for platforms to remove illegal content promptly and exploring how this mechanism balances the need for rapid response with the protection of freedom of expression. Further, the section analyses the DSA's adoption of a conditional liability framework for online intermediaries, contrasting it with a general monitoring obligation and discussing the implications for content moderation, innovation, and free expression. It illustrates the risks of automated content moderation systems required by the DSA through a hypothetical scenario, highlighting the potential for over-removal of legitimate content and broader implications for freedom of expression. Lastly, it examines the role of Digital Services Coordinators as specified in Article 38 of the DSA, detailing their responsibilities in monitoring compliance and approving 'trusted reporters,' and discussing potential concerns regarding the balance of power and judicial oversight. The chapter

²⁰⁴J Mchangama, N Alkiviadou and R Mendiratta (n 39) 7.

²⁰⁵J Barata (n 27) 12.

concludes by summarising the critical analysis of the DSA's provisions, emphasising the benefits and potential drawbacks, and stressing the need for policymakers to address concerns about over-removal of content and the balance of power.

3.2.1. Transparency Measures

According to the provisions of the DSA package all providers of server-based network-accessible services (Internet hosting providers) must justify their decisions on content moderation, set up an internal complaint-handling mechanism, and make use of out-of-court dispute resolution methods²⁰⁶. The DSA introduces some seemingly straightforward transparency measures²⁰⁷. For instance, the platforms have an obligation to inform their users of the reasons underlying the removal of content posted or shared by them on the platforms themselves. In practice, platforms are required to inform users when their content is removed for breaching community guidelines. This notification typically outlines the specific guideline violated (e.g., hate speech, graphic violence), details the content removed, and references the relevant section of the platform's terms of service or community standards. This transparent approach aids users in comprehending the reasons behind content removal and serves as an educational tool regarding unacceptable behaviour on the platform²⁰⁸. Transparency measures undoubtedly protect users' freedom of expression since interference should be justified by a precise overriding interest to be permitted. However, there is a risk of limiting individual autonomy

²⁰⁶J Mchangama, N Alkiviadou and R Mendiratta (n 39) 8.

²⁰⁷ The DSA introduces some seemingly straightforward transparency measures. Recital 3 of the DSA emphasises that "responsible and diligent behaviour on the part of intermediary service providers is essential for a safe, predictable and reliable online environment and to enable citizens of the Union and other persons to exercise their fundamental rights guaranteed by the Charter of Fundamental Rights of the European Union ('Charter'), in particular freedom of expression and information, freedom to conduct business and the right to non-discrimination." For a detailed overview, see the various contributions in B. Petkova and T. Ojanen (Eds.), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries*, Cheltenham-Northampton, 2020.

²⁰⁸ 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)

in expressing opinions, which may not always be justified by the need to protect other rights or values, like personal dignity, honesty, religious diversity, or different opinions²⁰⁹.

3.2.2 Systemic Risks in the DSA

At this stage, it is essential to examine the systemic risk delineated in the DSA. Systemic risks denote the potential unintended consequences or adverse effects that may arise from the implementation of regulatory frameworks such as the DSA. While the DSA addresses a broad range of systemic risks, this analysis focuses specifically on three main types of systemic risks that are particularly pertinent to the discussion of hate speech and content moderation. The first systemic risk pertains to the dissemination of illegal content. The regulation mandates that platforms must remove content deemed illegal under national and international laws upon receiving notice. This category includes, but is not limited to, hate speech, content related to terrorism, and material involving child sexual abuse. However, the regulation's broad and imprecise definition of 'illegal content' introduces significant interpretative ambiguities and challenges. This vagueness can result in the excessive removal of content, thereby potentially infringing upon and restricting freedom of expression. The second type of systemic risk "Impact on Fundamental Rights" involves the actual or foreseeable effects of the service on the exercise of fundamental rights protected by the Charter of Fundamental Rights of the EU (EU Charter), including human dignity, freedom of expression and information, privacy, data protection, non-discrimination, and the rights of the child. For example, the design of algorithmic systems used by very large online platforms can inadvertently exploit the weaknesses and inexperience of minors, leading to addictive behaviour or exposure to harmful content. The third type of systemic risk 'Manipulation and Public Safety Risks' addresses the risks of manipulation of very large online platforms and search engines, which can have adverse effects on democratic processes, civic debate, electoral processes, public safety, and public health. This includes the spread of disinformation, which can influence public opinion and electoral outcomes, as well as content that

²⁰⁹ J Mchangama N Alkiviadou and R Mendiratta (n 39) 8

could harm public health, such as misinformation about vaccines. These specific risks are selected for in-depth analysis because they directly relate to the core issues of hate speech and content moderation, which are central to this chapter. They represent significant areas where the DSA's impact on fundamental rights and freedom of expression can be critically assessed. Other systemic risks addressed by the DSA, though important, fall outside the primary focus of this discussion and are therefore not covered in this thesis. The regulation defines systemic risk as any negative impact on the exercise of fundamental rights, including respect for private and family life, freedom of expression and information, the prohibition of discrimination. According to Recital 79 and Art. 33 of the Digital Services Act (DSA), very large online platforms are identified as potential sources of systemic risks that could affect fundamental rights, public safety, and societal welfare. The recital mandates that these platforms undertake comprehensive assessments to identify and mitigate these risks. Key areas of focus include the dissemination of illegal content, the impact on fundamental rights such as freedom of expression, privacy, and non-discrimination, and the risk of manipulation that could compromise democratic processes, public health, and safety. It highlights the necessity for transparency and accountability in platform operations to maintain a secure and predictable online environment. This chapter will critically analyse these systemic risks within the framework of the Digital Services Act (DSA) to understand their potential impact on online content moderation. By examining the DSA's provisions, we can identify how the regulation addresses the dissemination of illegal content, its effects on fundamental rights, and the manipulation of platforms. This analysis is crucial for addressing the central research question: "How do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression"? By linking the systemic risks to the core aspects of the DSA, this chapter aims to evaluate whether the DSA effectively mitigates harmful content without excessively infringing upon freedom of expression and other fundamental rights. Additionally, understanding

these implications helps in assessing the practical challenges and potential unintended consequences that may arise from the implementation of the DSA. This critical examination provides insights into how regulatory frameworks can be improved to better safeguard both individual rights and public interests in the digital age.

3.2.3 Notice and Action Process in the DSA

The regulatory framework established by the EU incorporates a 'notice and action' process, detailed in Articles 16-18 of the DSA, to manage the removal of illegal content²¹⁰. Upon receiving such notices, hosting services are required to act without undue delay, considering both the nature of the illegal content in question and the urgency of the action required²¹¹. Discussing the 'notice and action' process is crucial because it highlights the procedural obligations imposed on platforms to remove illegal content swiftly. This mechanism is central to understanding how the DSA aims to balance the rapid response to harmful content with the protection of freedom of expression. By mandating prompt action, the DSA seeks to mitigate the potential harm caused by illegal content while ensuring that platforms do not arbitrarily or excessively censor user-generated content. This balance is critical to maintaining an open and free digital environment, where legitimate discourse is not unduly restricted. The importance of Article 16's 'notice and action' mechanism extends further when considered in conjunction with Article 34 of the DSA. Article 34 stipulates that all large online platforms conduct risk assessment to identify and evaluate significant systemic risks arising from the operation and utilization of their services within the EU. These assessments encompass various systemic risks, including the dissemination of illegal content, potential infringements on fundamental rights such as privacy, freedom of expression, and non-discrimination, and the intentional manipulation of services. The 'notice and action' mechanism is a critical tool for platforms to address one of the key systemic risks identified in these assessments: the dissemination of illegal content. By promptly responding to

²¹⁰ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 39) 16-18.

²¹¹ *Ibid.*, 16-18.

notices of illegal content, platforms can demonstrate compliance with the DSA's requirements and actively contribute to mitigating systemic risks. Moreover, the data and insights gained from the 'notice and action' process can inform the platforms' annual risk assessments, helping to identify patterns, emerging threats, and areas needing additional safeguards or policy adjustments. Thus, the interplay between Article 16 and Article 34 underscores a comprehensive regulatory approach. Article 16 provides the operational means to tackle illegal content in real-time, while Article 34 ensures that platforms maintain an ongoing, proactive stance in managing and mitigating broader systemic risks. Section 3.3 of this thesis will undertake an exhaustive examination of systemic risk within the regulatory ambit of the DSA, with particular emphasis on its compliance mechanisms and the implications for online intermediaries.

3.2.4 Conditional Liability Framework

The DSA has favoured a framework of conditional liability for online intermediaries over a general monitoring obligation. This framework of conditional liability is a key aspect of the DSA's approach, ensuring that intermediaries such as social media platforms or hosting services are held accountable for user-generated content only under specified circumstances. In contrast, a general monitoring obligation would require intermediaries to actively monitor and police all content on their platforms continuously. The adoption of conditional liability by the EU implies that intermediaries are not automatically held accountable for illegal content uploaded by users. Instead, their liability is contingent upon meeting certain conditions, such as the failure to promptly remove illegal content upon notification or the failure to implement sufficient measures to prevent its dissemination. This regulatory approach seeks to strike a balance between holding intermediaries responsible for unlawful content and preventing excessive burdens that could hinder online innovation and impede free expression²¹². Furthermore, it is essential to address the monitoring and compliance mechanisms specifically tailored for large platforms. The DSA mandates that only VLOPs are required to assess

²¹²J Barata (n 27) pp.3-24.

the systemic risks associated with their operations²¹³. This additional obligation is anticipated to produce similar effects concerning the attribution of liability to platforms, potentially increasing instances where operators might choose to limit their exposure to financial penalties by prioritising the removal of harmful material²¹⁴. In other words, to mitigate the potential costs associated with violations of the DSA and thus reduce the likelihood of sanctions²¹⁵, platform operators may feel compelled to engage in more extensive and thorough interventions than necessary. Such actions can interfere with users' freedom of expression²¹⁶.

3.2.5 Automated Content Moderation and Its Risks

The approach taken by platform operators towards over-intervention is particularly evident in the realm of automated content moderation. To illustrate the implications of the DSA, consider a scenario where a social media platform is mandated to promptly remove illegal content, such as hate speech, upon notification. To comply with these requirements and avoid financial penalties for non-compliance, the platform may deploy an automated content moderation system designed to rigorously flag and remove posts containing specific keywords commonly associated with hate speech. However, this approach carries the risk of inadvertently removing a significant amount of legitimate content that, while containing these keywords, is neither harmful nor illegal. For instance, it might delete a post discussing the detrimental effects of hate speech in an educational context or a news article reporting on incidents of hate speech. By intervening in such an overly comprehensive and stringent manner, the platform may inhibit users from freely discussing important issues, thereby

²¹³ See recital §79 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).

²¹⁴ J Barata (n 27) pp.3-24.

²¹⁵ *Ibid.*, pp.3-24.

²¹⁶ F. Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) European Centre for International Political Economy, pp.1-11.

encroaching on their freedom of expression. In this case, the platform's intervention is more penetrating and widespread than justified, as it exceeds the necessary steps to comply with the DSA, resulting in the removal of content that does not contravene any laws or guidelines²¹⁷. Moreover, the provision's wording seems to encompass both perspectives, suggesting that platforms might need to develop content moderation policies specifically targeting users, accounts, pages, etc., that have proven or are likely to become sources of illegal content²¹⁸. The most problematic aspect of this provision is the lack of clarity in the interpretation and application by platforms of several key elements, which are critical for the exercise of freedom of expression, such as the definitions of harmful content, the criteria for content moderation, and the mechanisms for appealing content removal decisions²¹⁹. This brings us to the next critical consideration: how platforms can balance the enforcement of the DSA with the protection of users' freedom of expression.

3.2.6 The Digital Services Coordinators in the DSA

To investigate the possible breaches that relate to content removal by platforms, the DSA envisages proposes the establishment of a Digital Services Coordinator in each member state to ensure its implementation in accordance with its own rules. This body will oversee investigating suspected breaches of obligations by VLOPs. The Digital Services Coordinators (Art. 49)²²⁰ must then approve 'trusted flaggers', whose reports must be processed by intermediaries 'in a timely manner²²¹'. The

²¹⁷ F Erixon (n 217)

²¹⁸ G Frosio & C Geiger (n 9) 31-77

²¹⁹ J Barata (n 28) 18

²²⁰ See Chapter 4, Section 1, Art. 49, 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).

²²¹ See Art. 22 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).

'trusted flaggers' must fulfil certain conditions, including objectivity and transparency of funding²²². Given that certain types of content—such as terrorist material or hate speech—are highly offensive and may require a certain level of technical expertise to be identified accurately and, above all, promptly, there has been a growing need over time to dedicate specific resources to their detection and removal from the web. This necessity has led to the emergence and proliferation of trusted flaggers, who are experts in the dynamics of content sharing on platforms and possess sufficient databases to develop effective tools for identifying illegal content. However, the incorporation of Digital Services Coordinator and 'trusted flaggers' into the DSA raises concerns. It creates a two-way removal system where private companies—non-state actors without judicial oversight—are directed to remove networked speech. This can be initiated by (1) public authorities, such as the Digital Services Coordinators, who have investigatory and enforcement powers to mandate content removal, and (2) trusted flaggers, whose reports are prioritised by platforms for swift action. Both mechanisms operate without direct judicial intervention, leading to potential risks of censorship or biased content moderation by non-judicial entities. Although these entities are intended to be independent, they can be influenced by state actors, which might compromise their neutrality. This two-way system raises concerns because it allows both state and non-state actors (including private companies and designated trusted flaggers) to influence content removal without direct judicial oversight. This setup can lead to potential abuses, such as censorship or biased content moderation influenced by political or commercial interests²²³.

3.2.7 Conclusion

In conclusion, although the DSA introduces significant and commendable measures to regulate online content and enhance transparency, it raises important concerns regarding its potential impact on

²²² *Ibid.*

²²³ Anand, R. and Bird, B. (2023). 'The Digital Services Act and Vulnerability: How Vulnerability is Defined by Digital Regulations in the EU'. *Irish Law Times*, 41(9), pp. 135-140. Available at: <https://doi.org/10.34961/researchrepository-ul.24064260.v1> (Accessed: 20 January 2024).

freedom of expression and the operational autonomy of online intermediaries. The DSA's focus on combating illegal content dissemination and its reliance on transparency measures are essential steps in fostering a safer digital environment. However, the broad definition of 'illegal content' and the conditional liability framework raises concerns about over-removal of content and potential stifling of legitimate discourse. Moreover, the introduction of Digital Services Coordinators and 'trusted reporters' poses challenges to the balance of power between private entities and state intervention, potentially undermining judicial oversight and exacerbating issues of censorship. Therefore, it is imperative for policymakers to address these critical concerns and ensure that the DSA strikes a balanced approach that effectively safeguards fundamental rights while upholding regulatory objectives in the digital sphere. Failure to do so risks compromising democratic values and perpetuating inequalities in online discourse.

3.3 The First Systemic Risk: Dissemination of Illegal Content – Defining and Addressing 'Illegal Content' in the DSA

The DSA addresses various categories of systemic risk, including the dissemination of illegal content, the impact on fundamental rights, and the risks of manipulation and public safety. The first category focuses on the dissemination of illegal content through online platforms²²⁴. An important issue within this category concerns the broad definition of 'illegal content' that platforms are required to remove upon notice²²⁵. Identifying illegal content is challenging due to varying and sometimes non-harmonized national laws across the EU, especially in categories like 'hate speech,' 'terrorism,' and

²²⁴ Recital 80, 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, <http://data.europa.eu/eli/reg/2022/2065/oj>.

²²⁵ Article 16 of the DSA specifically sets out the "notice and action" procedures that online platforms must implement. This includes the requirement for platforms to establish easy and effective ways for users to notify them about illegal content and to act promptly on these notifications. Additionally, Article 35 focuses on the responsibilities of very large online platforms to assess and mitigate systemic risks, which includes the dissemination of illegal content. *See* 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, <http://data.europa.eu/eli/reg/2022/2065/oj>.

'extremism,' which can lead to significant discrepancies and human rights concerns²²⁶. The regulation lacks guidance on addressing these complex and foreseeable challenges. Understanding the concept of illegal content is crucial because it forms the basis for content moderation policies and compliance mechanisms that platforms must implement. This approach is designed to foster a secure, predictable, and trustworthy online environment, thereby aligning with the primary objectives outlined in the Digital Services Act (DSA). Specifically, Article 1(1) of the DSA outlines the general objectives of the regulation, including the aim to ensure the proper functioning of the internal market for intermediary services, while Recital 79 emphasizes the importance of mitigating the dissemination of illegal content and disinformation to ensure a safer and more reliable digital space for users across the EU²²⁷. However, the ambiguity surrounding content removal requirements, coupled with their broad scope, could result in the excessive removal of content, thereby infringing upon users' rights to freely express their thoughts. By clearly defining what constitutes illegal content, the DSA seeks to balance the need to protect users from harmful material with the preservation of fundamental rights such as freedom of expression. On a practical level, mitigating the systemic risk of disseminating illegal content will probably require the use of artificial intelligence (AI)²²⁸. The systemic risk here involves the rapid spread of illegal content, such as hate speech or graphic violence, which can be difficult for human moderators to identify and remove swiftly due to the sheer volume of content generated online. AI presents a viable solution to mitigate this risk by employing advanced algorithms for content moderation. These AI-powered systems can analyse vast amounts of data quickly, identifying and removing unlawful content in real-time. By doing so, AI helps ensure compliance with DSA regulations, which mandate the prompt removal of illegal content to protect users and

²²⁶ J Barata (n 28) 18

²²⁷ 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, <http://data.europa.eu/eli/reg/2022/2065/oj>.

²²⁸ Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. *Regulating high-reach AI: On transparency directions in the Digital Services Act*. Internet Policy Review, 13(1), 2024, p. 13, <https://doi.org/10.14763/2024.1.1746>.

maintain a safe online environment. However, the use of AI raises complex issues such as transparency, bias, and user rights, which must be carefully managed to avoid infringing on freedom of expression and other fundamental rights. According to Söderlund et al. (2024), while AI can efficiently identify and remove unlawful content to comply with the DSA, its deployment raises issues of transparency, bias, and user rights²²⁹. The DSA mandates risk assessments and transparent procedures to ensure accountability, while independent audits help detect and rectify biases. A balanced approach combining AI and human oversight is essential to address complex decisions contextually. Protecting user rights through mechanisms for reporting and appealing content decisions is also critical. Ensuring ethical AI deployment in content moderation involves addressing these key considerations of transparency, fairness, and justice²³⁰. As a result, the deployment of AI systems for automated blocking, filtering, or removal of content may suffer from considerable harm. This harm occurs because existing pressures on internet companies to aggressively screen for illegal content, such as 'extremist', 'hate speech', or other unlawful content, worsen the problem of significant amounts of legally permissible content being removed according to the terms of service. Excessive and unreasonable burdens on the operators of online platforms could force them to implement monitoring systems capable of undermining the freedom to disclose information on the internet²³¹. This problem is directly linked to the systemic risk of disseminating illegal content, as overly broad or imprecise definitions of illegal content can lead to the over-removal of legitimate content, thus impacting users' freedom of expression. The lack of a clear and precise definition of what constitutes 'illegal content' raises several interpretative doubts as to what can be considered illegal content. This ambiguity complicates the task for AI systems as these systems rely on predefined criteria to identify and remove unlawful material. Therefore, in relation to the obligation for platform operators to inform their users, in the event of removal of content posted or shared by

²²⁹ Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling (n 229) 13

²³⁰ *Ibid.*, p.13

²³¹ J Barata (n 28) 17

them on the platforms (as previously discussed in section 3.2.3, the so-called ‘notice and action’ intervention) the underlying interference, impacting on the limitation of individual freedom of expression, is justified by the existence of specific illegal content²³². The DSA imposes not only the obligation to state why a particular content is considered unlawful, but also to adequately justify its notice, identifying the circumstances, context and nature of the alleged breach of law²³³. As previously discussed in section 3.2.3, the ‘notice and action’ mechanism is a regulatory process where platforms are required to remove illegal content upon notification. In implementing this mechanism, platforms often rely on algorithms to manage the vast amount of content that needs to be monitored. These technological tools are designed to predict and identify illegal content and, in some cases, anticipate where crimes are likely to occur. Law enforcement agencies may then use these predictions to prioritise police time for inquiries and arrests. However, the use of algorithms in this context can be problematic because it may lead to potential discrimination if the algorithms incorporate factors such as ethnicity and race²³⁴. For example, if the data used to train these predictive algorithms reflects existing societal biases—such as assumptions that certain crimes are more likely to be committed by specific segments of the population—the predictions might disproportionately target these groups. Algorithms can thus perpetuate existing discriminatory patterns by amplifying biases present in the data. Therefore, it is crucial to closely monitor and appropriately safeguard these systems to prevent such biases²³⁵. This discussion is relevant to the systemic risk of disseminating illegal content because the over-reliance on flawed predictive tools can lead to the unjust removal of legitimate content and the unfair targeting of specific groups, exacerbating the issue of discriminatory practices. The ‘notice and action’ mechanism, while intended to protect users from illegal material, can inadvertently cause harm by undermining users' freedom of expression and perpetuating systemic biases if not properly

²³²J Barata (n 28) 17

²³³*Ibid.*, p. 17

²³⁴ Mayson, Sandra G. "Bias in, bias out." *Yale LJ* 128 (2018) pp.2218-2300.

²³⁵*Ibid.*, pp.18-21

managed. Ensuring that these technological tools are transparent, fair, and just is essential to mitigating these risks and upholding the integrity of online discourse. Digital services are often built on pre-existing police databases, which may consciously or unconsciously reflect systematic bias. For instance, the inherent bias evident in the data collected during investigations, arrests, and stops, such as unfair profiling or crime prediction, even when the officers involved are not consciously discriminatory, leading to increased surveillance and enforcement actions against certain communities. As a result, this can lead to discriminatory outcomes on the grounds of ethnicity, race, or other factors triggering unequal access to services or other opportunities. Predictive algorithms may contribute to biased decision-making and discriminatory consequences, depending on how crimes are documented, which crimes are determined to be included in the research and which analytical techniques are employed²³⁶. Algorithms can learn to associate certain words, phrases, or topics with negativity or hate speech even when the context is not inherently hateful. Decisions made in community policing, particularly those influenced by prejudice, are likely to generate discussions or comment boards where existing biases are reinforced and strengthened. When incorporated into an automatic computer algorithm considered independent and unbiased, the police may not recognise bias, or prejudice based on a person's race or ethnicity²³⁷. Thus, bias is likely to grow more common and may not be questioned as such, making it harder to be discovered. We argue that the ambiguity surrounding content removal requirements, coupled with their broad scope, could result in the excessive removal of content, thereby infringing upon users' rights to freely express their thoughts. This ambiguity, when combined with algorithmic biases, may lead to the over-removal of content, particularly affecting marginalised groups or those expressing dissenting opinions. Even content that is not inherently hateful may be mistakenly interpreted as such by biased algorithms, resulting in its

²³⁶Mayson Sandra G (n 235) 18-21

²³⁷ *Ibid.*, pp.18-21

removal and potentially silencing legitimate voices. Considering the complexities outlined in this section, it becomes evident that while the DSA aims to address systemic risks associated with illegal content dissemination, its implementation poses significant challenges. The broad definition of 'illegal content' and the reliance on AI for content moderation raise concerns about potential over-removal of content, which could infringe upon users' freedom of expression. Additionally, inherent biases in algorithmic decision-making may disproportionately target certain groups and perpetuate discriminatory outcomes. These issues underscore the need for more precise definitions within the DSA, enhanced oversight of AI algorithms, and a critical evaluation of the regulatory framework's impact on online discourse and societal equality. If these issues are not addressed, there is a risk of stifling legitimate voices, undermining democratic principles, and perpetuating systemic inequalities in the digital sphere. Therefore, it is imperative for policymakers to rigorously scrutinise and revise the DSA to ensure a balanced approach that effectively tackles systemic risks while upholding fundamental rights and values in the online environment.

3.4 The Second Systemic Risk: The Actual or Foreseeable Negative Effects on the Exercise of Fundamental Rights

The second category of systemic risks regulated by the DSA concerns 'any negative effects for the exercise of fundamental rights²³⁸'. In the specific, it "concerns the actual or foreseeable effects of the service on the exercise of fundamental rights protected by the Charter, including, *inter alia*, human dignity, freedom of expression and information, including media freedom and pluralism, the right to privacy, data protection, the right to non-discrimination, and the rights of the child and consumer protection. Such risks may arise, for example, in relation to the design of the algorithmic systems used by the very large online platform or by the very large online search engine or the misuse of their service through the submission of abusive notices or other methods for silencing speech or hampering

²³⁸Recital 81, 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

competition. When assessing risks to children's rights, providers of very large online platforms and very large online search engines should consider, for example, how easy it is for children to understand the design and operation of the service and how they may be exposed through the service to content that could harm their health or their physical, mental and moral development. Such risks may arise, for example, in relation to the design of online interfaces that intentionally or unintentionally exploit the weaknesses and inexperience of minors or that may cause addictive behaviour²³⁹. They include fundamental rights content such as those associated with curtailing freedom of expression. This systemic risk is very important for our analysis for different reasons. The primary concern is that any infringement of a fundamental right is inherently unlawful²⁴⁰. The second issue pertains to the reference to 'any negative effects' in Recital 80 of the DSA, which is crucial in the context of human rights. This article ensures that very large online platforms assess and mitigate systemic risks, including those that impact fundamental rights, such as freedom of expression and privacy²⁴¹. For example, within the European framework - considering the broad context of the Council of Europe and the countries where the Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) is in force - the publication of news on matters of public interest can sometimes negatively affect to private and family life of certain public figures. However, this effect is usually subject to balancing in the light of the assessments inherent in the guarantee of other rights and freedoms - enshrined in the ECHR. This balancing act gives rise to protections defined by the case law of the European Court of Human Rights (ECtHR), establishing permissible limits, in a democratic society, to the right to freedom of expression and freedom of information²⁴². Similarly, the

²³⁹ Recital 81, 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

²⁴⁰ J Barata (n 27) 17.

²⁴¹ *Ibid.*, p. 18.

²⁴² In this regard, see the judgments of the Court of Justice of 11 September 2014, Case C-291/13, *Sotiris Papasavvas*, ECLI:EU:C:2014:2209; of 15 September 2016, Case C-484/14, *Tobias Mc Fadden*, ECLI:EU:C:2016:689.

protections afforded to minors against certain types of content must not be of such a nature as to deprive adults of safe access to the same type of information²⁴³. Based on this provision, national authorities could decide, for instance, that pressing criticism of public authorities themselves constitutes a systemic risk that needs to be addressed by online platforms²⁴⁴. This leads to an excessive exercise of control by national authorities. The third problem concerns the expression 'the foreseeable real effects of the service on the exercise of fundamental rights protected by the EU Charter of Fundamental Rights' contained in the same provision (Recital 81 of the DSA). Indeed, serious difficulties arise for the interpreter since this definition does not adequately consider either the seriousness of the violation or the frequent coexistence of several 'conflicting' rights²⁴⁵. Moreover, leaving the decision on how to carry out the balancing to the platforms themselves gives rise to serious concerns, since it seems unthinkable to require the platforms to articulate complex legal interpretations in relation to all the online content whose conflict with the rights to be protected is to be ascertained. Such interpretations must continue to be made, necessarily on a case-by-case basis, by national courts and the Strasbourg Court, once domestic remedies have been exhausted²⁴⁶. Therefore, the DSA proceeds to scale down the powers of platforms through the procedural harmonisation of content control, the establishment of ad hoc national and supranational public bodies and the conferral of specific enforcement powers on the Commission. In this way, the tendency to delegate public functions to the private companies owning the platforms is reduced, though not eliminated. To achieve a proper balancing of interests and rights, the legislator and the courts can build on a wealth of case law of the Court of Justice on various provisions of the e-commerce directive

²⁴³J Barata (n 27) 19

²⁴⁴J Barata (n 27) 19

²⁴⁵*Ibid.*, p.19

²⁴⁶*Ibid.*, p.19

and related acts, such as the provisions concerning the interpretation of the concept of ‘information society services’²⁴⁷ or those concerning the liability of intermediary service providers²⁴⁸.

3.5 Third and Fourth Categories of Systemic Risks: Manipulation of Very Large Online Platforms (VLOPs) and Very Large Search Engines (VLSEs) - Effects and Consequences

The third and fourth categories of systemic risks regulated by the DSA concern respectively 'actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes as well as public security' and the '[...] design, functioning or use, including through manipulation, of very large online platforms and of very large online search engines with an actual or foreseeable negative effect on the protection of public health, minors and serious negative consequences to a person's physical and mental well-being, or on gender-based violence'²⁴⁹. We analyse these two categories of systemic risks in parallel because the manipulation of VLOPs and VLSEs result in actual or foreseeable negative effects on democratic processes, civic debate and electoral processes, and public safety. These provisions are harbingers of serious consequences for the right to freedom of expression. One concern is the negative effects on health, electoral processes and public safety, as platforms may face legal responsibilities to restrict access to lawful content. Algorithm-based systems in companies like Meta might be better at removing biases but there is a risk of discrimination if

²⁴⁷ Judgments of the Court of Justice of 2 December 2010, Case C-108/09, *Ker-Optika bt*, ECLI:EU:C:2010:725; of 11 September 2014, Case C-291/13, *Sotiris Papasavvas*, ECLI:EU:C:2014:2209; of 15 September 2016, Case C-484/14, *Tobias Mc Fadden*, ECLI:EU:C:2016:689.

²⁴⁸ Cases C-236/08 to C-238/08 *Google France SARL and Google Inc. v. Google France SARL and Google Inc.*, ECLI:EU:C:2010:159; of 12 July 2011, Case C-324/09, *L'Oréal SA*, ECLI:EU:C:2011:474; of 24 November 2011, Case C-70/10, *Scarlet Extended*, ECLI:EU:C:2011:771; of 16 February 2012, Case C-360/10, *SABAM v. Netlog NV*, ECLI:EU:C:2012:85; of 27 March 2014, Case C-314/12, *UPC Telekabel Wien GmbH v. Constantin Film Verleih*, EU:C:2014:192.

²⁴⁹ Recital 82 and 83, 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

algorithms incorporate factors such as ethnicity and race²⁵⁰. Training data²⁵¹ reflecting societal biases can lead to predictions disproportionately targeting specific groups, thereby perpetuating and amplifying existing discriminatory patterns. It is crucial to monitor and safeguard these systems to prevent such biases. Algorithmic decision-making technologies may perpetuate or intensify indirect prejudice through stereotyping because they rely on the correlation between data sources and efficiency considerations²⁵². Indirect discrimination occurs when differentiated treatment cannot be justified. Therefore, avoiding unjustified algorithmic decision-making systems and building them properly is critical. Unjustified and illegal differential treatment arises when risk assessments are based on skewed data, which can result in indirectly discriminatory judgments. For example, using police statistics to develop risk profiles, such as the number of re-arrests, might be influenced by racial bias. Facebook users might see similar biases in algorithmic decision-making systems, which are harder to recognize and correct than human decisions. The broad and vague criteria for identifying 'harmful' content can lead to an overly broad interpretation influenced by differing political ideologies within the EU²⁵³. Consequently, platforms might reduce or limit borderline content, potentially damaging democracy and free debate, thus limiting freedom of expression²⁵⁴. The provisions examined adopt a 'negative' approach to speech, legitimising possible state interventions that would otherwise be unlawful if applied to other media. For instance, Germany's Network Enforcement Act (NetzDG) provides evidence of stricter content moderation laws²⁵⁵. Such provisions

²⁵⁰ Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling (n 229) 9

²⁵¹ Training data consists of datasets used to teach AI models, and if these datasets reflect societal biases, the AI can perpetuate and amplify those biases in its predictions and decisions. This can lead to unfair targeting of specific groups and reinforce existing discriminatory patterns. For further discussions on this topic See Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022), 35 pages. <https://doi.org/10.1145/3457607>

²⁵² Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling (n 229) 10

²⁵³ J Barata (n 28) 20

²⁵⁴ *Ibid.*, p.20

²⁵⁵ Wagner, Ben, et al. "Regulating transparency? Facebook, twitter and the German network enforcement act." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.

are problematic as they allow authorities to impose non-transparent restrictions based on unclear criteria, violating principles of clarity and predictability essential for the rule of law. This lack of clear legal frameworks creates an unpredictable environment, making it difficult for individuals to understand permissible speech and content removal criteria²⁵⁶. Moreover, the vague definition of these restrictions makes it impossible for authorities or courts to assess the necessity and proportionality of content removal or restrictions according to human rights standards. In our technological society, political, economic, and social life inherently involves potential dysfunctions and risks. It is crucial to evaluate how accurately Internet Intermediaries, contribute to ‘additional risks’ or elevate existing ones to an ‘unacceptable’ level²⁵⁷. Private platforms should not be solely responsible for such complex analyses due to potential biases and lack of expertise. Instead, careful consideration is needed to determine the best instruments to combat foreseeable negative effects. Emphasising the human rights repercussions that may arise from these tasks is essential.

3.6 Mitigation of systemic risks

Article 35 of the DSA examines methods to mitigate systemic risks, can be mitigated, including developing codes of conduct. The Commission and the Board will establish and support voluntary codes of conduct at the Union level to address systemic risks and illegal content, ensuring alignment with Union laws. These codes will have clear objectives, key performance indicators, and regular reporting and assessment mechanisms. In cases of non-compliance, corrective actions will be

²⁵⁶ T. MADIEGA, *Reform of the EU liability regime for online intermediaries. Background on the forthcoming Digital Services Act*, 2020, PE 649.404; N. IACOB, F. SIMONELLI, *How to Fully Reap the Benefits of the Internal Market for E-Commerce?*, 2020, PE 648.801; L. NIOMBO, T. EVAS, *Digital services act - European added value assessment*, EPRS PE 654.180; J. B. NORDEMANN *The functioning of the Internal Market for Digital Services: responsibilities and duties of care of providers of Digital Services*, 2020, PE 648.802; H. SCHULTE-NÖLKE ET AL., *The legal framework for e-commerce in the Internal Market- State of play, remaining obstacles to the free movement of digital services and ways to improve the current situation*, 2020, PE 652.707.

²⁵⁷J Barata (n 28) 20

requested to maintain the integrity and effectiveness of these code²⁵⁸. This article is crucial as it outlines the framework for managing systemic risks and illegal content on very large online platforms and search engines within the EU, emphasising the need for clear, transparent, and adaptable codes of conduct to protect public interests, ensure legal compliance, and enhance accountability among digital service providers. Measures taken to mitigate systemic risks must also be 'reasonable, proportionate and effective', as explicitly provided for in Article 35²⁵⁹. However, given the complexity of the tasks assigned to online platforms, these principles do not guarantee the necessary predictability regarding the measures and practices to be implemented. Furthermore, a platform's refusal to participate in the development and implementation of such a code of conduct, without adequate explanation, could be considered by the Commission when determining whether the platform has unreasonably breached its obligations. This compliance will be ensured through specific enhanced surveillance mechanisms explicitly dictated by the DSA framework²⁶⁰. Regarding risk

²⁵⁸ Art.45, 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

²⁵⁹ *Ibid.*

²⁶⁰ Pursuant to Article 50 of the DSA - 'Requirements for digital service coordinators:

1. Member States shall ensure that their Digital Services Coordinators perform their tasks under this Regulation in an impartial, transparent and timely manner. Member States shall ensure that their Digital Services Coordinators have all necessary resources to carry out their tasks, including sufficient technical, financial and human resources to adequately supervise all providers of intermediary services falling within their competence. Each Member State shall ensure that its Digital Services Coordinator has sufficient autonomy in managing its budget within the budget's overall limits, in order not to adversely affect the independence of the Digital Services Coordinator.
2. When carrying out their tasks and exercising their powers in accordance with this Regulation, the Digital Services Coordinators shall act with complete independence. They shall remain free from any external influence, whether direct or indirect, and shall neither seek nor take instructions from any other public authority or any private party.
3. Paragraph 2 of this Article is without prejudice to the tasks of Digital Services Coordinators within the system of supervision and enforcement provided for in this Regulation and the cooperation with other competent authorities in accordance with Article 49(2). Paragraph 2 of this Article shall not prevent the exercise of judicial review and shall also be without prejudice to proportionate accountability requirements regarding the general activities of the Digital Services Coordinators, such as financial expenditure or reporting to national parliaments, provided that those requirements do not undermine the achievement of the objectives of this Regulation.

mitigation measures, automatic filtering mechanisms may often be necessary to address systemic risks or to comply with the rules established by codes of conduct (Artt. 45 et seq.). Despite the DSA's transparency obligations regarding this mechanism, errors by automatic monitoring tools can seriously and irreversibly impact users' fundamental rights, including their privacy, freedom of expression and information, and protection against discrimination and due process violations²⁶¹. However, the DSA lacks clear and binding guidelines to guide the design and implementation of these measures, especially those that are potential sources of interference with human rights. Such measures include 'action orders' requiring platforms to remove specific illegal content, 'notice and action' mechanisms (as described in section 3.2.3) where users can flag potentially harmful content, and algorithmic transparency and systemic risks mitigation measures. The lack of relevant provisions necessitates that platforms and supervisory bodies adequately consider the human rights impact, particularly on freedom of expression, from implementing such mitigation measures²⁶². Algorithms are crucial for ensuring pluralism of information and transparency. However, as explained in section 3.3 of this chapter, these algorithms can have significant limitations, including biases and the potential for over-removal of legitimate content. Overly extensive content removals by platforms are problematic for two reasons: first, they stifle the freedom of speech of individual users who might want to discuss and debate the removed content; second, they set a negative precedent for social media platforms, making some generalised content removals appear necessary and proportionate. Thus, a focused approach to unify content removal and moderation practices is necessary to balance regulatory compliance with the protection of fundamental rights²⁶³. In conclusion, this thesis posits that the best framework to overcome these limitations – which include lack of clarity, potential biases,

²⁶¹Over the past two decades, Union legislation in the field has covered 'information society services', particularly through the E-Commerce Directive (2000) and the General Data Protection Regulation (GDPR, 2018). The DSA builds on these foundations by introducing comprehensive transparency obligations for automated content moderation.

²⁶² J Barata (n 28) 20

²⁶³ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 12

and over-removal of content- is provided by International Human Rights Law (IHRL). This framework offers comprehensive guidelines that can ensure measures are reasonable, proportionate, and effective while safeguarding fundamental rights. The next section will explore how IHRL can serve as a robust framework for the moderation of online content, ensuring that platforms operate transparently and fairly while protecting user's rights.

Part 2: International Law

3.7. International human rights law (IHRL) as a framework for the moderation of online content

Considering the analysis carried out so far, the DSA strengthens the obligations for platforms for VLPOs and VLSEs, by increasing their liability and imposing penalties for violations and delays. Additional fines not exceeding 1 percent of turnover, are imposed if platforms provide inaccurate information, fails to rectify such information, provides incomplete or misleading information by a staff member, or refuses to submit to an on-site inspection²⁶⁴. From the analysis of the DSA, it can be inferred that platforms will increasingly rely on automatic content moderation tools, incentivising the removal of more content rather than allowing it to remain visible. In this context, the legal standards for online content moderation, including criteria for the removal of illegal content, will be crucial to ensure the exercise of free and democratic speech²⁶⁵. Scholars advocate for adopting International Human Rights Law (IHRL) as the framework for moderating content on platforms. For example, Evelyn Aswad argues that the international law offers the most suitable legal framework to protect freedom of expression²⁶⁶. While platforms are not signatories to International Human Rights

²⁶⁴ Art. 52 (3), 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

²⁶⁵ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 12

²⁶⁶ E. M. ASWAD, 'The Future of Freedom of Expression Online' (2018) 17 Duke Law & Technology Review 1, pp. 52-53.

instruments, extending these norms to include private entities would enhance protection, albeit through voluntary adherence. Adopting a broad IHRL perspective - extending its application beyond state authorities to private platforms – could address legal challenges arising from the complex nature of the Internet. IHRL’s global orientation could help social media platforms operating in multiple jurisdictions navigate differing national laws and practices, promoting cohesive approach to online expression²⁶⁷. Moreover, employing IHRL based approaches could meet user demands for protection, as many users remain sceptical about state regulation of social media²⁶⁸. Adaptation to the IHRL would be extremely useful in moderating and regulating content falling within the realm of 'hate speech' that is widespread on online platforms in today's virtual, centralised, social media environment²⁶⁹. In conclusion, this thesis supports the adoption of the IHRL framework as the best available approach to guide content moderation practices. Due to the scope of this thesis, a deeper discussion on the comprehensive implementation of IHRL in content moderation cannot be provided here, and this remains a significant limitation of the current analysis.

3.8. Hate speech and International Human Rights Law

In regulating online hate speech, political and institutional bodies, as well as social media platforms, must be cautious of broadly worded bans on hate speech. Such bans can target individuals, associations, movements, or parties wishing to express dissenting views, including the groups that these restrictions aim to protect, such as ethnic and religious minorities²⁷⁰. To regulate online hate speech in a manner consistent with fundamental rights, platforms should define terms and conditions

²⁶⁷ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 12

²⁶⁸ *Ibid.*, p.12

²⁶⁹ *Ibid.*, p.12

²⁷⁰ T. Dias Oliva, D. M. Antonialli & A. Gomes, ‘Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online’ *Sexuality & Culture*, 702; Facebook’s Hate Speech Policies Censor Marginalized Users’ *Wired*. 2021 p. 25.

based on a precise threshold, established by Article 20(2)²⁷¹ of the International Covenant on Civil and Political Rights (ICCPR). The Rabat Plan of Action suggests a high threshold for defining restrictions on freedom of expression, incitement to hatred, and for the application of article 20 of the ICCPR. It outlines a six-part threshold test taking into account (1) the social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination and (6) likelihood of harm, including imminence²⁷². Article 20(2) of the ICCPR states that 'any propaganda of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence²⁷³ shall be prohibited by law'. The 2012 Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/67/357 (SRFOE) clarifies that the threshold of types of expression falling under Article 20(2) should be high and robust²⁷⁴. The Rabat Action Plan (RPA) further clarifies this threshold, setting a high bar for restricting freedom of expression. The RPA suggests assessing the seriousness of hate speech by considering the severity of the statements, the harm suffered, and the frequency, quantity, and extent of communications²⁷⁵. Restriction on freedom of expression must be the least intrusive measure, and it is crucial to emphasise that the right to freedom of expression under the IHRL admits "even expression that may be considered deeply offensive" and categories akin to

²⁷¹ United Nations. "International Covenant on Civil and Political Rights." Office of the High Commissioner for Human Rights, Article 20: "1. Any propaganda for war shall be prohibited by law; 2. Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law".

²⁷² Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22

²⁷³ General Comment No. 34 notes that Articles 19 and 20 are compatible and complement each other. The acts covered by Article 20 are all subject to restriction under Article 19(3). International Covenant on Civil and Political Rights.

²⁷⁴ Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/67/357 (7 September 2012), para. 45.

²⁷⁵ Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22

hate speech, such as denial of historical facts and religious insult or blasphemy²⁷⁶. Therefore, when dealing with hate speech, both policy makers and social media companies should consider the nature of hate speech as inferred from the texts mentioned above²⁷⁷. The use of the RPA allows for a practical assessment of the contested content, which can be carried out through the six-part test and focuses on evaluating profiles such as the context of reference and the likelihood of harm occurring, to ensure that the decision to remove the content from the platform corresponds to the parameters set by the IHRL²⁷⁸. Accordingly, platforms should ensure that their terms and guidelines for content moderation adequately reflect the standards set forth in Articles 19 and 20 of the ICCPR and, procedures to train those (natural persons, human beings) who use artificial intelligence²⁷⁹. If an IHRL approach were to be used to address online hate speech issues, the definition of hate speech currently applicable to Internet Service Providers (ISPs) could be significantly narrowed and thus raise the threshold for determining when deeply offensive and harmful speech can be removed. This approach, far from endorsing permissiveness, aims to balance conflicting rights and avoid excessive restrictions that may amount to censorship. Strengthening protection for controversial speech may have negative consequences, but these can be mitigated if platforms offer users tools to adopt personal filters or use third-party filters. This allows users to protect themselves from offensive content that does not meet the high threshold for hate speech under Article 20(2) of the ICCPR²⁸⁰. Our view is that while protecting freedom of expression is essential, it is equally important to ensure this freedom does not infringe on other fundamental rights, such as the right to be free from hate speech, incitement to violence, and disinformation. We advocate for a balanced approach that respects and protects all

²⁷⁶ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 14

²⁷⁷ See Articles 19 and 20(2) ICCPR, in HRC documents, SRFOE reports and the Rabat Action Plan (RPA)

²⁷⁸ J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 13-14

²⁷⁹ *Ibid.*, pp.13-14

²⁸⁰ *Ibid.*, pp.13-14

human rights, ensuring that content moderation practices are fair, transparent, and accountable. This perspective aligns with IHRL principles, providing a comprehensive framework for balancing rights and addressing the complex challenges posed by online content.

3.9 Conclusions

The potential threats to freedom of expression posed by the framework outlined in the Digital Services Act (DSA) demand careful consideration. By conferring authority upon states to regulate private entities—entities not necessarily governed by International Human Rights Law (IHRL)—the DSA enables expedited removal of content deemed illegal, enforced through the imposition of fines. Within DSA act that is binding all European Union (EU) member states, robust assurances of freedom of expression assume paramount significance. This holds particularly true in the contemporary context marked by a surge in censorship and authoritarian tendencies, wherein we find ourselves amid a period characterised by a perceived regression in the safeguarding of freedom of expression²⁸¹. Therefore, it is imperative for the EU to exercise prudence in avoiding granting unchecked discretion, particularly concerning sensitive facets like content moderation. It is crucial to recognise that platforms are dynamic entities, continuously evolving over time at the regulatory landscape. Notably, the DSA lacks provisions emphasising adherence to the doctrines and principles of International Human Rights Law (IHRL) in actions such as systemic risk assessments or content removal. Mandating private entities to swiftly and efficiently adjudicate the legitimacy of speech under the threat of fines, albeit contingent on reported content, poses inherent issues. Integrating IHRL provisions, particularly Article 19 and Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), into DSA provisions related to hate speech moderation is advocated. This integration would necessitate content moderation practices to adhere to the principles of legitimacy, necessity, and proportionality outlined in Article 19(3) of the ICCPR, thereby safeguarding freedom of expression. Establishing a clear definition of 'illegal content' under the DSA is imperative to

²⁸¹J.Mchangama, N. Alkiviadou and R. Mendiratta (n 42) 14

address hate speech effectively while upholding the rule of law. Additionally, attention to potential biases in algorithms used for tasks such as liability coverage, credit scoring, and differential pricing is warranted, as they may perpetuate discrimination. Discriminatory biases in algorithms can exacerbate existing prejudices, necessitating measures to mitigate such risks and ensure fair treatment. This thesis posits that applying IHRL principles to the design and implementation of these algorithms can provide a robust framework for addressing these biases. IHRL emphasises equality and non-discrimination, which are essential in developing fair and unbiased algorithmic systems. By integrating IHRL standards, platforms can better ensure that their content moderation practices do not perpetuate or amplify existing social inequalities, thus promoting more equitable digital environment.

Chapter 4: Regulatory Approaches to Online Harm. A Comparative Analysis of European and English Legislation

Summary: 4.1 Introduction - 4.2 The scope and limitations of the Online Safety Act: Balancing Regulation and Innovation in Platform Governance - 4.3 Revisiting Online Harm: Doctrinal Debates on Online Safety Regulatory Approaches - 4.4 Navigating the complexity of harmful content regulation in the Online Safety Act – 4.5 Platform accountability: a comparison between the ‘Duty of Care’ in the Online Safety Act and the ‘Good Samaritan’ Rule in the Digital Services Act – 4.6 Conclusions

4.1 Introduction

This chapter explores the legal frameworks in the United Kingdom (UK) that govern the liability of internet intermediaries in the context of regulating online hate speech, while also safeguarding the right to freedom of expression. The moderation strategies employed by social media platforms, including Instagram, Facebook, TikTok, and YouTube, which host user-generated content, often lack transparency, resulting in widespread distrust²⁸². To address these concerns, the UK government has implemented legislation, such as the Online Safety Act (OSA), to ensure responsible behaviour from both users and platforms. Similarly, the European Union (EU) has introduced the Digital Services Act (DSA) to achieve comparable objectives. Setting a threshold for what constitutes ‘harm’ has historically played a crucial role in regulatory intervention²⁸³, but defining ‘harm’, especially online, remains complex²⁸⁴. This debate was evident when the UK proposed broadening the definition of “harm” to include not only illegal content but also legal but harmful content. The initial draft of the

²⁸²Bartolo, Louisa & Matamoros Fernandez, Ariadna. “Online Harm”. *ISP-Platform Governance Terminologies Essay Series*, 27 June 2023, p.1-16.

²⁸³Persak, Nina. “Criminalising harmful conduct: the harm principle, its limits and continental counterparts”. Springer Science & Business Media (2007), pp.1-160.

²⁸⁴ Park, Tae Jung, and Akshita Rohatgi. "Balancing the platform responsibility paradox: A case for amplification regulation to mitigate the spread of harmful but legal content online." *Computer Law & Security Review* 52 (2024): 105960, pp.1-16.

UK's Online Safety Bill (OSB) faced significant criticism for including 'legal but harmful' content, leading to fears of government censorship of lawful speech. The revised bill, which was enacted into law in October 2023, eliminated obligations concerning such content. For content outside the scope of criminality but still harmful, such as promoting eating disorder, requires platforms to provide users with tools such as content filters, reporting mechanism, and customisable content preference to manage and avoid such content. This chapter argues for broadening the definition of 'online harm' in safety regulations to encourage diverse solutions to tackle lawful harms. The discussions around the OSA and its controversies prompt broader questions about the suitability of legal definitions of 'harm' for all forms of online harm. It is crucial to consider whose viewpoints might be marginalised when specific notions of 'harm' are prioritised, particularly those related to psychological and emotional well-being, as well as threats to public safety and security. This chapter examines key elements of the OSA's structure and compares it with the DSA, focusing on the practical difficulties of implementing these distinctions into legislation. The OSA and the DSA, while sharing common goals of regulating online content to protect users, diverge significantly in their scope, definitions, responsibilities, enforcement mechanisms, and approaches to user empowerment and transparency²⁸⁵. The scope and coverage of the OSA are notably broad, targeting a wide array of online harms that encompass both illegal content and harmful but legal content. This multifaceted approach mandates platforms to protect users from diverse harmful activities. In contrast, the DSA, while also addressing illegal content, places a stronger emphasis on transparency and accountability measures for online platforms. It aims to create a safer digital space through stringent rules on content moderation, advertising, and platform responsibilities, yet does not mandate the same level of proactive measures against legal but harmful content as initially proposed by the OSB. Defining harm in the context of online regulation presents a complex challenge²⁸⁶. The OSB originally included provisions for 'legal but harmful'

²⁸⁵Beatriz Kira and Laura Schertel Mendes, 'A Primer on the UK Online Safety Act' (Verfassungsblog 2023), pp.1-5. Available at: <<https://verfassungsblog.de/a-primer-on-the-uk-online-safety-act/>> (Accessed on 14 November 2023).

²⁸⁶ *Ibid.*, pp.3-5

content, sparking concerns over potential overreach and censorship. This led to revisions in the final version, which now emphasises illegal content and mandates tools for users to manage exposure to harmful content. Conversely, the DSA maintains a clearer distinction between illegal and harmful content but focuses regulatory efforts primarily on the former. It incorporates measures to tackle disinformation and protect minors without mandating the proactive mitigation of legal but harmful content as initially proposed by the OSB²⁸⁷. The responsibilities imposed on platforms also differ between the two acts. The OSA institutes a ‘duty of care’, compelling platforms to take reasonable steps to shield users from harm. This includes specific obligations to prevent the dissemination of illegal content and to mitigate risks associated with harmful content²⁸⁸. On the other hand, the DSA establishes a tiered framework of responsibilities based on the platform's size and impact. While all platforms must adhere to transparency and reporting obligations, larger platforms are subject to additional requirements such as risk assessments and independent audits. Regulatory enforcement under the OSA is centralised with the UK’s Ofcom, which has the authority to issue fines, impose sanctions, and mandate improvements in platform practices. The DSA, however, designates national authorities within each EU member state to enforce its regulations, with the European Commission playing a coordinating role, especially concerning very large online platforms (VLOPs)²⁸⁹. This decentralised enforcement model aims to leverage local oversight while ensuring a unified approach across the EU. User empowerment and transparency are critical components of both regulatory frameworks, yet they approach these aspects differently. The OSA mandates that platforms provide users with tools to manage and avoid harmful content, thereby emphasising user empowerment. Additionally, it requires platforms to maintain transparency about their terms of service and content moderation policies. The DSA also prioritises transparency, obligating platforms to clearly communicate the reasons behind content removal or restriction and to offer user complaints and

²⁸⁷ B Kira and L Schertel Mendes (n 285) 3-5

²⁸⁸ *Ibid.*, pp.3-5

²⁸⁹ *Ibid.*, pp.3-5

dispute resolution mechanisms. However, the DSA's focus extends further into systemic risks posed by algorithms, particularly for large platforms, which must conduct risk assessments and implement measures to mitigate these risks. In summary, the distinctions between the OSA and the DSA lie in their respective approaches to scope, definition of harm, platform responsibilities, regulatory enforcement, and user empowerment. These differences underscore the practical challenges of harmonising online safety legislation while balancing the need for effective regulation with the protection of fundamental rights and innovation in digital spaces. This chapter begins by introducing the focus on regulatory approaches to online harm, providing an overview of the OSA. It emphasises the importance of balancing regulation, innovation, and free speech in platform governance. The chapter examines the scope and limitations of the OSA, discussing its provisions and objectives. It highlights the reliance on future codes of conduct from Ofcom and critiques the Act's limited impact on broader content moderation practices. Following this, the chapter revisits the concept of online harm through doctrinal debates, exploring its definition in Anglo-American criminal law and liberal theories. It discusses the cumulative and social harms associated with online content, arguing for a societal perspective in online safety regulations, and critiques the individualistic approach in the OSA. Subsequently, the chapter navigates the complexity of harmful content regulation in the Online Safety Act, examining definitions and regulatory measures, the challenges of content moderation, and the potential for over-censorship. It critiques the exclusion of 'legal but harmful' content from the OSA. The chapter then offers a comparative analysis of the OSA's duty of care model and the DSA's Good Samaritan rule, exploring how each regulatory framework addresses platform responsibilities for content moderation. It evaluates the practical implications for platforms and the effectiveness of these approaches in promoting safer online environments. In conclusion, the chapter summarises the key findings from the comparative analysis, reflects on the effectiveness of current regulatory approaches in addressing online harm, discusses the need for continuous adaptation of regulations to keep pace with technological advancements, and offers final thoughts on balancing regulation, innovation, and the protection of fundamental rights in digital spaces. This analysis aims to achieve the objective of

understanding how the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression, providing insights into the practical challenges and effectiveness of these regulatory approaches.

4.2 The scope and limitations of the Online Safety Act: Balancing Regulation and Innovation in Platform Governance'

The Online Safety Act (OSA) received Royal Assent on October 26, 2023, signifying a relevant step in UK platform regulation. Its goal is to promote careful and diligent content management. The Act broadly defines illegal content as anything that constitutes an offence, covering various forms such as words, images, speech, or sounds that are offensive or harmful²⁹⁰. Moreover, the Act includes "prioritised illegal content," such as terrorism-related or child abuse content, mandating platforms to take measures to minimise its presence²⁹¹. Additionally, the OSA introduces new speech-related offences, including false communication, aiming to prevent dissemination of false information causing harm. It imposes new responsibilities on tech companies regarding illegal online content like child exploitation, fraud, and terrorism, aligning the UK with other jurisdictions tightening online safety rules. Historically, determining a threshold for "harm" has been pivotal in defining the boundaries of regulatory intervention²⁹². The concept of "harm," especially in the context of online activities, is complex and lacks a universally accepted definition among regulatory bodies and digital platforms²⁹³. This issue surfaced prominently when the UK proposed a broad definition of "harm" in

²⁹⁰UK Government. (2023). *Online Safety Act: explainer*. GOV.UK. Available at: <https://www.gov.uk/government/publications/online-safety-act-2023-explainer> (Accessed on 24 January 2024).

²⁹¹*Ibid.*

²⁹²L Bartolo and A Matamoros Fernandez (n 285) 6-12

²⁹³Victoria Nash, *Revise and Resubmit? Reviewing the 2019 Online Harms White Paper*, 11 J. MEDIA L. 18, 22 (2019), pp.1-8.

its online safety regulations²⁹⁴, addressing not only illegal content and behaviour, such as terrorism and incitement to violence, but also content harmful to children. The Act mandates that platforms take measures to prevent children from accessing harmful content, such as content promoting self-harm, even if it is not illegal. The OSA in the UK represents a legislative endeavour aimed at fostering a safer digital environment by holding internet companies accountable for the content circulated through their platforms²⁹⁵. The regulatory approach involves categorising responsibilities based on the legality of content, effectively establishing a framework that scrutinises content for compliance. As part of this framework, the Act mandates that platforms take measures to prevent children from accessing harmful content, such as content promoting self-harm, even if it is not illegal. Under the OSA, internet entities are mandated to adhere to a statutory duty of care, monitored by Ofcom, the United Kingdom's independent communications regulator, which is set to take on regulatory powers to protect UK users from harms occurring on user- to-user services and search services²⁹⁶. This statutory framework encompasses various responsibilities applicable to operators facilitating user-to-user interactions and providing search services, thereby ensuring a comprehensive oversight mechanism for online safety within the UK's digital landscape. However, the Online Safety Act (OSA) faces significant limitations due to its lack of detailed guidance on the specific actions required from social media platforms and search engines, relying instead on future codes of conduct from Ofcom and pending secondary legislation²⁹⁷. The specific measures that platforms must undertake will be determined by future codes of conduct and guidelines to be issued by Ofcom, the designated regulator, along with forthcoming secondary legislation. The absence of precise details exacerbates

²⁹⁴See Press Release, Nadine Dorries, Secretary of State, UK Dep't for Digit., Culture, Media & Sport, Statement Made on 7 July 2022 [Statement UIN HCWS194] (July 7, 2022).

²⁹⁵ B Kira and A Schertel Mendes (n 285) 3.

²⁹⁶ Lorna Woods, 'The Duty of Care in the Online Harms White Paper', 11 *Journal of Media Law* 6, 2019, pp.6-17.

²⁹⁷ B Kira and A Schertel Mendes (n 285) 5.

the complexity of implementation, creating uncertainty about how platforms should comply with the regulations²⁹⁸. The legislative approach places strong emphasis on finer points yet to be established, making the intricacies of implementation even more critical compared to similar European regulations. The OSA is characterised by ambiguity in its definitions of key terms and categories, particularly those relating to illegal content and content harmful to children. As above mentioned, illegal content is defined as ‘content that amounts to a relevant offence’²⁹⁹, covering a range of situations from the use of harmful words, images, or sounds, to the possession, viewing, or accessing of such content. As already explained, the full operational timeline of the OSA remains uncertain due to the pending finalization and presentation of harm mitigation codes of conduct. While initial drafts have been published and are under consultation, the complete codes of practice need to be laid before Parliament, leaving the timeline for full operation unclear. This situation creates significant implementation challenges for platforms attempting to adhere to the Act's requirements, as they must navigate a landscape of evolving guidelines and secondary legislation. A fundamental concern regarding the OSA is whether it will significantly impact content moderation practices beyond illegal content. The OSA primarily targets illegal content and content harmful to children, with minimal influence on broader content moderation practices. However, these provisions were removed following criticism and concerns about free speech and potential government overreach. Consequently, the Act's impact on overall content moderation remains limited. The OSA imposes broad mandates on platforms without clearly defining thresholds or criteria for categorising platforms or delineating their responsibilities. This vagueness could inadvertently create an intermediary liability regime, where platforms face legal repercussions for failing to meet ambiguous standards. Such a regime could place undue burden on platforms, complicating their compliance efforts and potentially stifling innovation. Additionally, the OSA's approach to regulating AI and emerging

²⁹⁸ B Kira and A Schertel Mendes (n 285) 5.

²⁹⁹ GOV.UK. (2023). “*Online Safety Act 2023*.” Available at: <https://www.legislation.gov.uk/ukpga/2023/50/enacted>. (Accessed on 24 January 2024).

technologies is underdeveloped³⁰⁰. The Act does not explicitly address the challenges posed by AI-powered chatbots and other advanced technologies, leading to uncertainty about whether these services fall under its purview³⁰¹. This gap in the legislation raises questions about the OSA's ability to adapt to rapid technological advancements and effectively regulate new forms of digital communication. Contrary to the DSA (examined in the third chapter of this thesis), which clearly defines the scope of regulation, including the specific duties of platforms related to content moderation, user protection, and transparency requirements, the OSA broadly describes service duties without specifying which platform features fall under regulatory scrutiny, such as content monitoring³⁰². The issue with leaving matters undecided and ambiguous lies in the potential for unlawful discretion by platforms or the regulatory authority. Our main concern in the context of online safety regulations is the potential for codes of conduct issued by regulatory bodies, such as Ofcom, to place undue burdens on service providers by holding them accountable for activities beyond their control. For instance, if Ofcom requires social media platforms to swiftly identify and remove content that are considered hate speech within 24 hours, it could impose excessive burdens due to the vast volume of content they host. This would necessitate substantial resources for effective monitoring and moderation, particularly straining smaller platforms. Moreover, the ambiguity and context-dependency of hate speech complicate the task of accurate and timely content assessment. Determining whether content constitutes hate speech often requires expert analysis and a nuanced understanding of cultural and social contexts, as well as the intent behind the speech, which cannot always be accurately evaluated within a short timeframe. Consequently, overly stringent requirements may prompt platforms to err on the side of caution, potentially leading to the removal of legitimate content and stifling free speech. This scenario highlights the delicate balance required in formulating

³⁰⁰B Kira and A Schertel Mendes (n 285) 4-5.

³⁰¹B Kira and A Schertel Mendes (n 285) 4-5.

³⁰²Trengove, Markus, et al. "A critical review of the Online Safety Bill." *Patterns* 3.8, 2022, p.23.

such regulatory codes to avoid unintended consequences, such as the suppression of valuable discussions and public discourse. Further complicating matters, imposing strict legal liabilities on service providers for non-compliance with these codes could expose them to numerous lawsuits and financial penalties. This legal risk is particularly problematic when decisions to remove content are subjective or when content lies in a grey area. Platforms might face legal challenges both from users whose content was erroneously removed and from those claiming insufficient action against harmful content. An example of these dynamics can be seen during the COVID-19 pandemic, where platforms faced the challenge of curbing misinformation about the virus and vaccines³⁰³. Regulatory requirements for the immediate removal of potentially misleading information might have led to the suppression of legitimate scientific debates about vaccine efficacy and safety, illustrating the potential for regulatory overreach. Similarly, in the context of hate speech, stringent legal frameworks could result in the suppression of free expression if platforms over-censor content to avoid liability, highlighting the delicate balance between regulating harmful speech and protecting freedom of expression. Supporting this argument, Balkin³⁰⁴ discusses the challenges faced by platforms in balancing content moderation with the need for transparency and fairness in their enforcement mechanisms. He highlights that overly burdensome regulations can lead to unintended consequences, such as the suppression of legitimate speech and disproportionate operational burdens on platforms, particularly smaller ones³⁰⁵. This echoes a similar provision in the DSA, where regulators may establish codes of conduct, or guidelines based on best practices³⁰⁶. To clearly and firmly establish

³⁰³ Angel N Desai, Diandra Ruidera, Julie M Steinbrink, Bruno Granwehr, Dong Heun Lee, Misinformation and Disinformation: The Potential Disadvantages of Social Media in Infectious Disease and How to Combat Them, *Clinical Infectious Diseases*, Volume 74, Issue Supplement_3, 15 May 2022, Pages e34–e39, <https://doi.org/10.1093/cid/ciac109>.

³⁰⁴ Balkin, J. M. (2018). *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*. UC Davis Law Review, 51, pp.1149-1210.

³⁰⁵ *Ibid.*, pp.1149-1210.

³⁰⁶ Art.45 of the European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Official Journal of the European Union.

their commitment to focusing on procedural elements while addressing concerns about censorship in the Online Safety Act, Ofcom, Parliament, and the Department for Digital, Culture, Media, and Sport (DCMS) should take several steps. In a recent *memorandum*, the DCMS stated that the Codes of Practice will emphasise the procedural aspects of user-to-user and search services³⁰⁷. However, to address worries about censorship targeting specific content, it's crucial for Ofcom, Parliament, and DCMS to solidify this commitment. Ofcom should draft and publish specific Codes of Practice, detailing clear guidelines for platforms. Parliament should review and approve these codes, incorporating public consultations to gather diverse input and ensure fairness³⁰⁸. The DCMS should conduct transparent consultations to refine the guidelines based on stakeholder feedback. Regular reviews and updates by Ofcom will keep the guidelines current, while clear public communications from all three bodies will reinforce their commitment. Additionally, Parliament should be ready to amend the legislation if needed. These steps will collectively ensure the effective implementation of the OSA while preventing undue censorship. Clarity is crucial, especially given that the Online Safety Act is less rigorous and precise compared to jurisdictions with content-specific regulations. In this context, it is imperative that Ofcom drafts specific Codes of Practice, which should then be reviewed and approved by Parliament. These Codes of Practice must also be refined through transparent consultations conducted by the DCMS to ensure they effectively address the nuances of online safety. These steps are essential for establishing clear guidelines for platforms and ensuring the effective implementation of the legislation. Additionally, the issue of clarity and specificity will be further explored in the subsequent section, which presents a comparative analysis of the OSA and the DSA.

³⁰⁷ Press Release, UK Department for Digital, Culture, Media & Sport & the Rt Hon Michelle Donelan MP, “New Protections for Children and Free Speech Added to Internet Laws” (Nov. 28, 2022). Available at: <https://www.gov.uk/government/news/new-protections-for-children-and-free-speech-added-to-internet-laws> (Accessed on 15 March 2024)

³⁰⁸ *Ibid.*

In this section, we have detailed the evolution, scope, and regulatory mechanisms of the Online Safety Act (OSA), emphasising the need for specific Codes of Practice by Ofcom, reviewed and refined through parliamentary oversight and public consultation, to ensure clarity and effective implementation, and compared these aspects with the more precisely defined Digital Services Act (DSA). This section argues that while the government's concerns about the risks of the Internet are legitimate, especially for vulnerable groups, it is crucial to maintain an open Internet that fosters free speech and innovation. Balancing these priorities requires trade-offs between security, freedom, and innovation, which the State must be able to manage. The subsequent paragraph will revisit doctrinal debates on online safety regulatory approaches, specifically focusing on the concept of online harm.

4.3 Revisiting Online Harm: Doctrinal Debates on Online Safety Regulatory Approaches

This section examines the concept of harm to better understand the progress and gaps in online safety regulation within the OSA. Concerns arose in late 2021 when the UK's Online Safety Bill (OSB) focused solely on individual harms like cyberbullying and harassment³⁰⁹. The approach of the Online Safety Act (OSA) includes focusing on the responsibilities and obligations of individual platforms based on their specific characteristics and functions. Instead of applying blanket regulations to all platforms, the OSA differentiates duties of care based on factors such as the size, reach, and nature of the services provided by each platform. This approach acknowledges that different platforms pose varying levels of risk to users and society and seeks to address these risks accordingly. For example, the OSA categorises platforms into different groups (Category 1, Category 2a, and Category 2b) based on their characteristics, with each category subject to different levels of regulatory obligations³¹⁰. We will analyse this categorization in detail in paragraph 4.5 of this chapter. The

³⁰⁹ P.Coe (n 23) 50-75

³¹⁰ GOV.UK. (2023). "*Online Safety Act 2023*." Available at: <https://www.legislation.gov.uk/ukpga/2023/50/enacted> (Accessed on 13 March 2024).

approach adopted in the OSA aims to ensure that platforms are held accountable for the harms associated with their services while also recognising the diverse nature of online platforms and the content they host. The individualistic approach, reflected in both the OSB and the Online Safety Act (OSA), classifies platforms based on their characteristics and imposes specific obligations to address these direct harms. When examining the UK's OSA, it's important to differentiate between individual and societal harms to fully grasp the progress and shortcomings in online safety regulation. Individual harms, such as cyberbullying and harassment, directly affect personal well-being and are often managed through user-level tools and platform-specific measures such as blocking and reporting functions, content filters, privacy settings, and dedicated support resources provided by the platform. However, this narrow focus overlooks broader societal harms like misinformation, hate speech and systemic abuse, which demand a more comprehensive regulatory perspective³¹¹. A societal harms lens, in contrast, could address these broader issues by recognising that social media platforms are part of complex ecosystems with significant societal interdependencies³¹². Addressing social harms necessitates a unique approach to validation, one that goes beyond individual experiences to understand the broader societal implications of online content and activities. The OSA acknowledges this need by emphasising the importance of considering the collective impact of harmful online behaviour when formulating regulatory measures and enforcement strategies. To enhance the regulatory framework, the OSA could adopt strategies like those used in environmental law. Environmental regulations have established a robust framework for accountability, transparency, and compliance, which can be adapted to enhance online safety regulations³¹³. For example, implementing mandatory impact assessments for new technologies or significant changes to online

³¹¹L Bartolo and A Matamoros Fernandez (n 285) 8-11.

³¹²Sally Broughton Micova, "*What Is the Harm in Size?*" CTR. ON REGUL. EUR.", Oct. 19, 2021, p. 6, <https://cerre.eu/publications/what-is-the-harm-in-size>.

³¹³Smuha, Nathalie A. "Beyond the individual: governing AI's societal harm." *Internet Policy Review* 10.3, 2021, pp.1-32.

platforms could help foresee and address potential societal harms. Public oversight mechanisms would ensure that platforms comply with safety regulations, while procedural rights would empower the public to access information, participate in decision-making processes, and seek justice for societal harms caused by online activities. By leveraging these regulatory strategies, the OSA can strengthen its effectiveness in safeguarding public welfare in the digital age, addressing both individual and societal impacts comprehensively. Nathalie Smuha suggests that environmental law offers valuable insights, as regulators have devised ‘societal mechanisms’ to address and prevent widespread environmental damage, such as pollution³¹⁴. These mechanisms involve public scrutiny to enhance accountability, mandatory impact assessments considering social impact, and citizens' rights to access information without needing to justify it based on individual harm risks³¹⁵. Smuha’s suggestion underscores the potential of adapting these well-established regulatory frameworks to tackle societal challenges posed by emerging technologies and their impacts on public welfare. In the context of the OSA, these insights from environmental law can be particularly valuable. The OSA aims to regulate digital platforms to prevent and mitigate societal harms, such as misinformation, cyberbullying, hate speech and harmful content. Connecting this to the research question of this thesis, “how do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression”, it is evident that shifting from an individualistic approach to a societal harm’s perspective aligns with broader academic discussions on harm. In Anglo-American criminal law, figures like Mill³¹⁶ and Feinberg³¹⁷ highlight the importance of distinguishing between genuinely harmful actions and those that are

³¹⁴Smuha Nathalie A (n 313) 12.

³¹⁵*Ibid.*, p.12.

³¹⁶ JOHN STUART MILL, ON LIBERTY (1859); Ten, Chin L., ed. *Mill's on liberty: A critical guide*. Cambridge University Press, 2008, p.1-106.

³¹⁷ Feinberg, Joel. “*The moral limits of the criminal law: volume 3: harm to self.*” Oxford University Press, USA, 1989.

merely unpleasant³¹⁸. Applying this distinction to online safety, just as Mill and Feinberg emphasise freedom of speech unless it directly incites violence or defamation, online safety regulations need to differentiate between individual harms and cumulative societal harms. This broader perspective ensures that regulatory frameworks address the systemic nature of online abuse and misinformation, promoting a safer and more equitable digital environment³¹⁹. My research thus investigates how these legal frameworks balance the liability of internet intermediaries with the right to freedom of expression, considering both individual and societal harms. This comprehensive approach is crucial for effective regulation in the digital age. However, scholars in the field of socio-legal studies have emphasised the drawbacks of this approach, particularly regarding its effectiveness in addressing everyday instances of abuse that may not individually cross legal boundaries but collectively have significant impacts on marginalised communities over time³²⁰. Cumulative harm refers to the situation in which an individual experiences harm due to the repeated occurrence of multiple negative effects, even though each effect may not individually meet the criteria for harm under criminal law frameworks³²¹. It is the recurrence of these negative experiences and their interconnectedness, intensifying one another as they accumulate, that renders them collectively harmful³²². Understanding the concept of cumulative harm is valuable when considering how to regulate online safety. This is particularly important because many online abuses are widespread, and those most affected are often from marginalised communities. Moreover, certain aspects of online platforms can contribute to the

³¹⁸ L Bartolo and A Matamoros Fernandez (n 282) 5.

³¹⁹ L Bartolo and A Matamoros Fernandez (n 282) 6.

³²⁰ Galván, Nicolás Quaid. "Adopting the Cumulative Harm Framework to Address Second-Generation Discrimination." *Colum. J. Race & L.* 11 2021, pp. 147-194.

³²¹ L Bartolo and A Matamoros Fernandez (n 285) 6.

³²² Friedlaender, Christina. "On microaggressions: Cumulative harm and individual responsibility." *Hypatia* 33.1 (2018), pp. 1-21.

accumulation of harm³²³. Studies have demonstrated that instances of men's violence against women on digital platforms often take the form of frequent, minor forms of abuse³²⁴. These actions can leave women feeling uneasy, uncomfortable, or insecure³²⁵. However, online harm typically doesn't arise as isolated incidents or fleeting actions. Rather, content and behaviour online often leave lasting traces that are easily accessible, searchable, and capable of being reproduced³²⁶. Opponents of traditional views on harm argue that relying heavily on individual-focused criminal law frameworks neglects the broader societal harms, which are equally important in addressing online harm regulation³²⁷. The impact of a 'harm', like racism, sexism, or repeated offences, extends beyond just one person and affects a broader community³²⁸. Therefore, the idea of cumulative harm is closely linked with that of social harm³²⁹. For instance, acts of 'casual' racism, such as jokes³³⁰ or stereotyping, contribute to cumulative harm because they occur within a context of systemic racial oppression³³¹. The unchecked accumulation of harm perpetuates social oppression, particularly in the case of racial

³²³Hoffmann, Anna Lauren, and Anne Jonas. "Recasting justice for Internet and online industry research ethics." *Internet Research Ethics for the Social Age: New Cases and Challenges*. M. Zimmer and K. Kinder-Kuranda (Eds.), np Bern, Switzerland: Peter Lang, Forthcoming, 2016.

³²⁴Jane, Emma A. "Misogyny online: A short (and brutish) history." (2016): 1-152.

³²⁵Gillett, Rosalie. "This is not a nice safe space": investigating women's safety work on Tinder." *Feminist Media Studies* 23.1 (2023), pp. 199-215.

³²⁶Boyd, Danah. "Social network sites as networked publics: Affordances, dynamics, and implications." *A networked self*. Routledge, 2010, (Zizi Papacharissi ed., 2010), pp. 1-18.

³²⁷Hillyard, Paddy, and Steve Tombs. "From 'crime' to social harm?" *Crime, law and social change* 48 (2007), pp. 9-25.

³²⁸Perry, Barbara. "Exploring the community impacts of hate crime." *The Routledge international handbook on hate crime*. Routledge, 2014.

³²⁹ Bryce, India, et al. "When quantity takes on a quality of its own: A retrospective exploration of the lived experience of cumulative harm." *Child abuse review* 32.4 (2023), pp. 1-10.

³³⁰Levchak, Charisse C., and Charisse C. Levchak. "Microaggressions, macroaggressions, and modern racism." *Microaggressions and modern racism: Endurance and evolution* (2018).

³³¹ Strand, Palma Joy, and Claudia E. Cohen. "Microaggressions and macro-injustices: How everyday interactions reinforce and perpetuate social systems of dominance and oppression." *Understanding and Dismantling Privilege* 12.2 (2022), pp.1-30.

discrimination³³². Essentially, it implies that if negative experiences and injustices are not addressed or corrected over time, they can compound and reinforce existing inequalities and injustices in society, particularly those related to racial discrimination³³³. Social harms can accumulate gradually over time, spanning extended periods, and involve intricate webs of causation³³⁴. Consequently, when online safety regulations primarily address individual harms, there's a danger of overlooking the broader societal impact of online activities³³⁵. This section argues that a social harm perspective in online safety laws would broaden the scope to address racial and gender-based abuse, acknowledging its wider societal impact beyond individual targets, contrary to the individualistic approach contained in the OSA. In this paragraph, we critically examine the OSA by distinguishing between individual and societal harms, reflecting on the limitations and progress in the regulatory framework. I highlight that the OSA and its predecessor, the Online Safety Bill (OSB), predominantly focus on individual harms such as cyberbullying and harassment, which are managed through user-level tools and platform-specific measures like blocking and reporting functions, content filters, and privacy settings. However, this narrow focus neglects broader societal harms, including misinformation, hate speech, and systemic abuse, which require a more comprehensive regulatory perspective. We argue that addressing societal harms necessitates a shift to recognising the complex interdependencies of social media ecosystems and the cumulative nature of harm. Drawing parallels with environmental law, we propose adopting similar regulatory strategies, such as mandatory impact assessments and public oversight mechanisms, to enhance accountability, transparency, and compliance. This perspective aligns with broader academic discussions on harm, notably the distinction made by Mill and Feinberg between genuinely harmful actions and those that are merely unpleasant. By integrating the concept of cumulative harm, especially as it affects marginalised communities, we emphasise the need for

³³² *Ibid.*, pp.1-3

³³³ Stand Palma J and Cohen E. C. (n 331) 1-3

³³⁴ Smuha (n 313) 4-6.

³³⁵ Willard, Nancy E. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press, 2007.

regulations to address both individual and societal impacts comprehensively. This approach is crucial for creating a safer and more equitable digital environment, balancing the liability of internet intermediaries with the right to freedom of expression. Additionally, we critique the traditional individual-focused frameworks for neglecting the broader societal impacts, arguing for a regulatory approach that considers the systemic nature of online abuse and misinformation. This comprehensive examination highlights the necessity of addressing both immediate and long-term harms in the digital landscape. With this foundation established, the next section, "Navigating the complexity of harmful content regulation in the Online Safety Act," will delve into the intricate challenges and strategies involved in effectively regulating harmful content within the framework of the OSA. This includes exploring detailed mechanisms and policy considerations essential for addressing both individual and societal harms in the digital age.

4.4 Navigating the complexity of harmful content regulation in the Online Safety Act

The previous section revisited the concept of online harm by exploring various doctrinal approaches. The aim of this section is to analyse the concept of harmful content as defined in the Online Safety Act (OSA). This analysis seeks to understand not only what the legislation means by harmful content, but also the implications when such harmful content is disseminated online. The OSA lacks specificity in defining harmful content and measures beyond content removal. This ambiguity can lead to dangerous over-censorship. The moderation process is widely distrusted due to its lack of transparency. As discussed earlier, platforms frequently face criticism for either removing too much content or not removing enough, deleting inappropriate content inconsistently, or not recognising how their systems promote harmful behaviour³³⁶. Consequently, both national governments like the UK and the EU have taken steps to make online platforms responsible for user-generated content and behaviour considered harmful. These actions, discussed in chapter three of this thesis, include

³³⁶Kölbel, Tobias, and Daniel Kunz. "Mechanisms of intermediary platforms." *arXiv preprint arXiv:2005.02111*, 2020, pp. 1-19.

methods like 'notice and action', risk evaluations, transparency initiatives, and setting up audit standards, often enforced with the possibility of fines. However, the absence of a specific damage threshold renders the existing legislation ineffective. Defining 'harm', especially in the online environment, continues to be a complex and ambiguous task for both regulators and the platforms under their jurisdiction. This ambiguity poses challenges in creating clear, enforceable guidelines and standards for what constitutes harmful content.³³⁷ The addition of a classification for 'lawful but harmful content' in the initial draft of the UK Online Safety Bill, which disappeared in the Act has faced significant backlash due to concerns about its potential to enable government censorship of lawful and constitutionally protected speech. In January 2023, a modified version of the Bill was reintroduced to Parliament and has since been enacted into law³³⁸. In the revised version, obligations concerning content categorised as 'legal but harmful' and accessible to adults have been eliminated from the law. Platforms are no longer required to evaluate the risk related to certain types of 'legal but harmful' content and conduct on their platforms³³⁹. Certain content previously categorised as "legal but harmful" may now be deemed illegal, following the government's declaration of plans to criminalise actions like promoting self-harm and violence, as well as sharing intimate images of individuals without their consent³⁴⁰. For content that doesn't meet the criteria for criminal relevance or fall under the definition of 'harmful' as outlined in the Bill and now in the Act, such as promoting eating disorders, online platforms must introduce extra tools for users to control their exposure to

³³⁷ V Nash (n 293), 1-8.

³³⁸ GOV.UK. (2023). "*Online Safety Act 2023*." Available at: [Legislation.gov.uk. https://www.legislation.gov.uk/ukpga/2023/50/enacted](https://www.legislation.gov.uk/ukpga/2023/50/enacted) (Accessed in 24 January 2024).

³³⁹ The final version of the Online Safety Bill, the Online Safety Act, has undergone a significant shift. Unlike earlier drafts that aimed to tackle a wider spectrum of harms, both illegal and legal but undesirable, the current iteration focuses solely on illegal content. Previously, the Bill envisioned a dynamic list of these "lawful harms" established by Parliament, requiring large platforms to assess risks and outline how they would address them in their terms of service. This broader approach has been scrapped in favour of a more limited focus on illegal activity.

³⁴⁰ V Nash (n 293), 1-8.

such content and aid in individual prevention efforts³⁴¹. Platforms must enforce their own Terms of Service (ToS), giving them the authority to decide which lawful harms, if any, to address, and emphasising content moderation over systemic strategies³⁴². Platforms are required to enforce their own ToS, which means they have the authority to decide which types of lawful but harmful content to address. This gives platforms significant discretion in determining their approach to content moderation. By enforcing ToS, platforms can remove or restrict content that, while not illegal, violates their internal rules. This includes hate speech, misinformation, and content that promotes self-harm or eating disorders³⁴³. However, the exclusion of provisions addressing ‘legal but harmful’ content in the new Bill under the Online Safety Act is viewed as a missed opportunity to enhance regulatory measures. This omission is perceived as a significant gap, failing to fully leverage the potential to protect users from a broader spectrum of harmful online behaviours. Embracing a wider perspective on the notion of online harm within online safety regulation is crucial. It involves moving beyond conventional legal frameworks and inspiring platforms to adopt a varied and balanced set of solutions to tackle permissible harm on their platforms³⁴⁴. While there is a definite concern about excessive regulation by governments linked to expansive interpretations of harm, we contend that sticking to narrow legal perspectives on online harm in state regulations doesn't sufficiently safeguard users. The meanings of "harm" are not rigid or predetermined, indicating the importance of scrutinising who has the authority to define ‘harm’ in emerging online safety rules and platform guidelines³⁴⁵. This chapter posits that marginalised or disregarded perspectives should be included when prioritising certain definitions of harm and specific evidentiary standards over others. It

³⁴¹*Ibid.*, pp.1-8.

³⁴²Johnson, Mark. "Save Online Speech Coalition-Online Safety Bill Committee Stage amendment briefing" 2022.

³⁴³Johnson, Mark (n 342)

³⁴⁴Goldman, Eric. "Content moderation remedies." *Mich. Tech. L. Rev.* 28, 2021, pp.1-59.

³⁴⁵Jiang, Jialun Aaron, et al. "Understanding international perceptions of the severity of harmful content online." *PloS one*, 2021.

contends that the inclusivity of diverse viewpoints is essential for a comprehensive and equitable understanding of harm³⁴⁶. The situation in the UK demonstrates a widespread misinterpretation regarding the consequences of enacting legislation that encourages platforms to enhance their moderation of lawful yet harmful speech and actions. Yet a regulatory framework centred on ‘systems and processes’, initially suggested in the UK and incorporated to tackle ‘systemic risks’ in the EU’s DSA, might necessitate platforms to provide risk evaluations and openly disclose their methodologies for addressing different kinds of lawful harms. These could encompass harms exacerbated by the platforms' technical functionalities, like seamless sharing or algorithmic amplification³⁴⁷. This section revisited the concept of online harm, particularly within the context of the UK’s OSA. It analyses the definition and implications of harmful content as outlined in the OSA, highlighting the Act's ambiguity and potential for over-censorship due to a lack of specificity in defining harmful content and beyond content removal measures. The discussion extends to the broader legislative landscape, including criticisms of the 'legal but harmful' classification and the challenges posed by vague definitions of harm. The section underscored the need for platforms to adopt varied and balanced solutions to manage permissible harm and the importance of inclusive definitions that consider marginalised perspectives. It also noted the omission of provisions addressing "legal but harmful" content as a missed opportunity for enhanced regulation and protection against a wide spectrum of harmful online behaviours. This analysis has paved the way for a deeper exploration of platform accountability. The subsequent section will compare the 'duty of care' obligations in the OSA with

³⁴⁶Matias, Nathan, Eric Pennington, and Zenobia Chan. "Testing Concerns about Technology's Behavioural Impacts with N-of-one Trials." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp.1722-1732.

³⁴⁷In the UK, it appears that the revised Online Safety Act will uphold the risk assessment responsibilities of platforms about behaviour and content that is judged to be a "danger" to minors. It appears that platforms will no longer be required to generate risk assessments about content, conduct, and systems that could endanger adults. See UK Government. (2023). *New protections for children and free speech added to internet laws* [Press release].

the 'Good Samaritan' rule in the EU's DSA. By examining how these two regulatory frameworks approach platform accountability and the mitigation of online harms, the discussion will highlight differences and potential synergies between the UK's and the EU's strategies. This comparison will shed light on how these frameworks seek to balance user protection with the principles of free expression and platform innovation.

4.5 Platform accountability: a comparison between the 'duty of care' in the OSA and the 'good Samaritan' rule in the DSA

In this section, the analysis focuses on the comparison between the OSA and the DSA in relation to hate speech and content moderation. This section reveals how the OSA introduces a model of 'Duty of Care', shifting from intermediary liability to proactive prevention of harm. As already mentioned in the second chapter of this thesis, in the EU, internet platform liability was managed by the e-Commerce Directive from 2000. Under this framework, certain types of internet services—like conduits (which merely pass information along), caching providers (which store temporary copies), and hosting platforms (which provide space for user content)—were protected from being legally responsible for the content they handled, but only if they didn't know it was illegal. Essentially, these platforms were shielded from legal trouble if they didn't know about any illegal content passing through them. However, once they did find out about illegal content, they were required to promptly remove it. This approach is commonly referred to as 'notice-and-takedown', as previously discussed in section 2.9. These legal protections were specifically outlined for platforms offering 'mere conduit', 'caching', or 'hosting' services, as detailed in Articles 12-14 of the directive³⁴⁸, as previously discussed in section 2.3. In the UK, online service providers were shielded from legal responsibility for what users posted on their platforms, as they adhered to provisions like the EU's E-

³⁴⁸Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the internal market," [2000] OJ L 178/1 ('ECD').

Commerce Directive. After Brexit, the UK is no longer bound to follow the EU's rules³⁴⁹. Nonetheless, the United Kingdom now prioritises the regulation of social media platforms, predominantly managed by private entities. These platforms decided what is allowed and what is not, making their own rules. Over time, these rules became more complex, turning into detailed systems and processes utilised by social media networks to control content based on their policies³⁵⁰. However, these self-regulated rules have not kept pace with the significant changes in online interactions and the profound impact of social media on society, prompting calls for new regulations. The OSA suggests a new approach to regulating online platforms, which was outlined in the White Paper on online harm and discussed during a Parliamentary subcommittee meeting³⁵¹. The approach to regulating online platforms outlined in the White Paper on Online Harms was carried forward into the OSA. Both documents emphasise a statutory duty of care requiring platforms to protect users, particularly vulnerable groups, from various online harms. This includes conducting risk assessments and ensuring transparency³⁵². The White Paper also proposed appointing an independent regulator to enforce these regulations, a provision that was adopted in the Online Safety Act. Instead of holding platforms responsible only if they knowingly let harmful content stay up (known as 'intermediary liability'), the OSA proposes a 'Duty of Care' model³⁵³. This means platforms would be held accountable if they consistently fail to take necessary precautions, as outlined in the law, to keep users safe. It is about making sure platforms actively work to prevent harm rather than merely reacting to

³⁴⁹Kira, Beatriz. "Pro-Competition and Online Safety: A More Holistic Approach to Digital Platform Regulation." *Research Handbook on Competition Law and Data Privacy*, Edward Elgar Publishing, 2023, pp. 1-19.

³⁵⁰B Kira (n 349) 1-19.

³⁵¹Neudert, Lisa-Maria. "Regulatory capacity capture: The United Kingdom's online safety regime." *Internet Policy Review*, 2023, pp 1-33.

³⁵²B Kira (n 349) 1-19.

³⁵³Park, Tae Jung, and Akshita Rohatgi. "Balancing the platform responsibility paradox: A case for amplification regulation to mitigate the spread of harmful but legal content online." *Computer Law & Security Review* 52, 2024, pp.1-16.

it afterward. The 'Duty of Care' concept holds platforms accountable only for their own deeds. This implies that if they fail to take appropriate action to stop or address issues, such as unlawful content or negative user behaviour, they may be held liable. Although they are not directly responsible for users' actions, they are accountable for their responses to those actions or outcomes, such as neglecting to remove harmful content when necessary³⁵⁴. In the EU, internet intermediaries that made a sincere effort to handle improper or illegal content were not directly protected by the e-Commerce Directive. However, the Digital Services Act (DSA) goes beyond this by introducing the so-called 'Good Samaritan' protection principle³⁵⁵. Article 7 of the DSA permits service providers to undertake voluntary actions to investigate, detect, identify, and remove illegal content. However, there are concerns that such proactive measures might lead to platforms being deemed as playing an active role, thereby gaining knowledge or control over the content and potentially forfeiting their liability exemption. In practice, conducting these investigations does not automatically render platforms liable for the content³⁵⁶. To address these concerns and provide clarity, the Good Samaritan clause was incorporated into the DSA, ensuring that platforms can voluntarily remove illegal content without jeopardizing their liability exemption. This principle states that if intermediaries voluntarily investigate or take measures to take down unlawful material, or if they abide by national and EU laws, they will not lose their liability's exemptions for hosting services³⁵⁷. It is important to note that this 'Good Samaritan' protection clause only applies if the intermediary acts honestly and responsibly. Recital 26 of the DSA and Article 7 emphasise balancing the rights of service providers and internet

³⁵⁴Dias, T. "Amending Online Safety Bill to Ensure Consistency with Core International Human Rights Instruments: Specific Recommendations.", 2022, pp.1-14.

³⁵⁵G'sell, Florence, The Digital Services Act (DSA): A General Assessment (April 15, 2023). in Antje von Ungern-Sternberg (ed.), Content Regulation in the European Union – The Digital Services Act, TRIER STUDIES ON DIGITAL LAW, Volume 1, Verein für Recht und Digitalisierung e.V., Institute for Digital Law (IRDT), Trier April 2023, Available at SSRN: <https://ssrn.com/abstract=4403433> or <http://dx.doi.org/10.2139/ssrn.4403433>

³⁵⁶ *Ibid.*, p.7

³⁵⁷Buri, I., & Van Hoboken, J. (2021). *The Digital Services Act (DSA) proposal: A critical overview*. Digital Services Act (DSA) Observatory, Institute for Information Law (IViR), University of Amsterdam.

users³⁵⁸. Stakeholders value the DSA codification of the ‘Good Samaritan’ norm because it allows them to defend themselves by demonstrating that they took steps to detect infringing content. The DSA incorporates principles akin to the ‘Good Samaritan’ principle, providing legal protections to online intermediaries that voluntarily remove or restrict access to harmful or illegal content³⁵⁹. This framework encourages proactive content management without compromising liability exemptions. The DSA mandates due diligence measures, such as implementing notice-and-action mechanisms, conducting risk assessments, and enhancing transparency, to manage risks and protect users. By balancing due diligence with liability protection, the DSA ensures that platforms can moderate content in good faith without being considered publishers, thus maintaining their liability exemptions while promoting a safer online environment³⁶⁰. The rule upholds a recent ruling made by the CJEU in the YouTube case,³⁶¹ which clarified that taking measures to detect infringing content (like copyright violations) does not automatically mean the intermediary has knowledge or control over all uploaded content. The concluding section of Article 7 of the DSA guarantees that intermediaries will retain their liability exemptions, even if they are required to undertake extensive due diligence to adhere to EU or national regulations, including the stipulations within the DSA itself. This is to avoid a scenario where complying with legal requirements makes them overly in the content they host which could otherwise strip them of their liability exemptions.³⁶² In practice, these due diligence requirements include monitoring for illegal content, ensuring transparency in their operations,

³⁵⁸ Schwemer, Sebastian Felix. "Digital Services Act: a reform of the E-Commerce Directive and much more." *Research handbook on EU internet law*. Edward Elgar Publishing, 2023, pp. 232-252.

³⁵⁹ Peguera, M. The Platform Neutrality Conundrum and the Digital Services Act. *IIC* 53, 2022, pp. 681–684. <https://doi.org/10.1007/s40319-022-01205-7>

³⁶⁰ Peguera, M. The Platform Neutrality Conundrum and the Digital Services Act. *IIC* 53, 2022, pp. 681–684. <https://doi.org/10.1007/s40319-022-01205-7>

³⁶¹ Joined Cases C-682/18 and C-683/18, *Frank Peterson v. Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH (C-682/18) and Elsevier Inc. v. Cyando AG (C-683/18)*, EU:C:2021:503.”

³⁶² Wilman, Folkert. "The EU's System of Knowledge-Based Liability for Hosting Services Providers in Respect of Illegal User Content-Between the e-Commerce Directive and the Digital Services Act." *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12 (2021): 317-341.

conducting risk assessments, and implementing measures to mitigate identified risks. By meeting these requirements, intermediaries must engage with the content to a certain extent, such as by removing illegal content or acting on user reports. However, Article 7 ensures that such involvement does not disqualify them from the liability protections provided by the DSA, thereby striking a balance between compliance with legal obligations and maintaining their role as neutral intermediaries³⁶³. In the UK, the ‘Duty of Care’ outlined in the OSA is influenced by principles from tort law, which deals with civil wrongs causing harm. This duty, central to tort law, requires individuals or entities to reasonably ensure they do not harm others. The Act draws from both general principles and specific duties seen offline, like those for property owners and employers who must ensure safety³⁶⁴. The concept of ‘Duty of Care’ legally obligates online platforms to take proactive and reasonable measures to prevent or mitigate known harms resulting from the content they host. This duty requires platforms to engage actively in managing risks associated with harmful content, such as cyberbullying, hate speech, and misinformation, by implementing content moderation, transparency reporting, and conducting risk assessments. The objective is to safeguard users from foreseeable harms while ensuring a balanced approach to content regulation³⁶⁵. These duties require proactive measures to prevent foreseeable harm, like safety protocols or warnings. Failure to fulfil these duties can lead to legal liability. The nature of the platform and the dangers associated determine how much of this obligation is carried out. Duty bearers may be compelled by law to take reasonable steps to protect users within foreseeable risks online³⁶⁶. While the ‘Duty of Care’ mandates proactive

³⁶³ *Ibid.*

³⁶⁴ Dias, T. “*Amending Online Safety Bill to Ensure Consistency with Core International Human Rights Instruments: Specific Recommendations.*”, 2022, pp.1-14. <https://www.elac.ox.ac.uk/wp-content/uploads/2022/03/hatespeechandtheonlinesafetybill-supplementaryevidencesubmission-diaspdf.pdf>

³⁶⁵ Kira, Beatriz; Schertel Mendes, Laura: “*A Primer on the UK Online Safety Act: Key aspects of the new law and its road to implementation*”, VerfBlog, 2023.

³⁶⁶ Woods, Lorna. “Regulating to Minimise Harm to Children and Young People.” *Children, Young People and Online Harms: Conceptualisations, Experiences and Responses*. Cham: Springer International Publishing, 2024.

risk management and user protection measures as a legal requirement, the 'Good Samaritan' rule encourages voluntary content moderation by offering liability protection for good faith actions. Both concepts aim to enhance online safety but operate through different legal and regulatory mechanisms. Online platforms have a duty of care to take reasonable steps to prevent or mitigate known negative impacts of speech, without removing certain content or portions of it to fully prevent those effects. It is about taking appropriate steps to address problems rather than always removing content entirely³⁶⁷. Defining 'online harm' clearly can alleviate concerns about vagueness and prediction in the online world. Platforms should evaluate and share how their recommendation algorithms affect users to better understand and tackle potential harms³⁶⁸. Regardless of whether the Parliament officially calls these responsibilities a 'Duty of Care', the core duties of platforms should align with the idea of human rights due diligence. This implies that businesses, whether they are located offline or online, have a social responsibility to consider how their actions may impact human rights.³⁶⁹ The OSA adopts a 'due diligence'³⁷⁰ approach, imposing obligations on platforms to actively manage online harm risks. However, these obligations vary based on the platform's size and the content it hosts. The law categorises platforms into groups based on their risk level. Category 1 platforms like Facebook, Twitter, and Instagram, which have high risk and impact, face the strictest obligations. Category 2a includes expansive search services like Google and Bing, while category 2b covers services with potential risks such as user-to-user services like Reddit or smaller online forums³⁷¹. The Act also

³⁶⁷Leiser, Mark, and Dr Edina Harbinja. "Content not available: Why The United Kingdom's Proposal For A "Package Of Platform Safety Measures" Will Harm Free Speech." *Technology and Regulation* 2019 (2020), pp.78-90.

³⁶⁸ Cobbe, Jennifer, and Jatinder Singh. "Regulating recommending motivations, considerations, and principles." *Considerations, and Principles*, 2019, pp. 1-49.

³⁶⁹M Leiser (n 367) 78-90.

³⁷⁰The 'due diligence' approach in the context of the Online Safety Act (OSA) refers to the set of proactive responsibilities imposed on online platforms to manage and mitigate risks associated with harmful content on their services. This approach requires platforms to take reasonable steps to prevent, identify, and address online harms, thereby ensuring a safer online environment for users. See UK Government. (2023). *Online Safety Act: explainer*. GOV.UK. Retrieved from <https://www.gov.uk/government/publications/online-safety-act-2023-explainer> (Accessed

³⁷¹*Ibid.*

targets providers of adult content, mandating measures to prevent minors' access. Unlike EU laws, the OSA doesn't set explicit criteria for each category, leaving it to the UK government and Ofcom to determine. Furthermore, the requirements vary according to the type of content, with a focus on content that is illegal and harmful to minors. Platforms must assess these risks and take appropriate actions. Moreover, user-to-user platforms must promptly remove illegal content upon detection. Thus, OSA obligations hinge on platform size, content assessment, and potential child exposure³⁷². As already mentioned in the section 3.5, the EU Digital Services Act introduces a risk-based system where companies are required to meet different levels of due diligence depending on the type of services they offer whether they're general intermediary services, online platforms, or classified as very large online platforms (VLOPs) or very large online search engines (VLOSEs)³⁷³. The DSA imposes significant responsibilities on very large online search engines (VLOSEs) and very large online platforms (VLOPs). These obligations cover various aspects of their operations, including content moderation policies, enforcement methods, transparency in information sharing, and how they make money from content, like online advertising³⁷⁴. Platforms must also address systemic risks to public health fundamental rights, political processes and individual well-being. Essentially, the DSA aims to make platforms accountable for the systems and processes they use to manage content, rather than just individual pieces of content³⁷⁵. The 'Duty of Care' outlined in the OSA is influenced by principles from tort law, which deals with civil wrongs causing harm. This duty requires individuals or entities to reasonably ensure they don't harm others. The Act draws from both general principles and specific duties seen offline, like those for property owners and employers who must

³⁷²See UK Government. (2023). *Online Safety Act: explainer*. GOV.UK. Retrieved from <https://www.gov.uk/government/publications/online-safety-act-2023-explainer>.

³⁷³For a more detailed examination see Cauffman, Caroline, and Catalina Goanta. "A new order: the digital services act and consumer protection." *European Journal of Risk Regulation* 12.4 (2021), pp.758-774.

³⁷⁴For an overview of the structure of the DSA, Buri, Ilaria, and Joris van Hoboken. "The Digital Services Act (DSA) proposal: a critical overview." *Digital Services Act (DSA) Observatory* (2021).

³⁷⁵Buiten, Miriam C. "The Digital Services Act from Intermediary Liability to Platform Regulation." *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12, 2021, pp.1-26.

ensure safety. The OSA has a narrower scope compared to the DSA. The differences can be categorised based on scope and focus. In terms of scope, DSA has a broader focus as it tackles illegal content, protects fundamental user rights, and even includes consumer law protections for online transactions. It aims to create a safer and fairer online environment overall. OSA on the other hand focuses on illegal content and the obligations online platforms must tackle it, thus the scope is more limited. The OSA now primarily addresses illegal content and content harmful to minors³⁷⁶. The legislation categorises obligations based on content legality, establishing a content-based regulatory framework³⁷⁷. Internet companies are subjected to a statutory duty of care overseen by Ofcom, the telecommunications regulator³⁷⁸. The OSA assigns varying responsibilities to operators, including those offering 'user-to-user services' and 'search services'³⁷⁹. All platforms falling under the law's scope must address illegal content and content harmful to children, promptly removing specified priority illegal content upon notification³⁸⁰. Like the DSA, companies must conduct risk assessments to identify potential risks and implement mitigation measures. However, unlike EU law, the UK legislation does not mandate identifying 'systemic risks' but rather focuses on the risk of users encountering illegal content or children being exposed to harmful content. This means that the EU focuses on identifying and addressing the root causes of online harms, such as algorithms that consistently promote hateful content while UK legislation tackles specific instances of harmful content thus failing to address reasons behind harmful content. In this section, we conduct a detailed

³⁷⁶ For an analysis of the OSA's structure and obligations, see Neudert, Lisa-Maria. "Regulatory capacity capture: The United Kingdom's online safety regime." *Internet Policy Review* 12.4 (2023), pp.1-34.

³⁷⁷ Michalis, M. 2024. Competition Law and Regulation. in: Puppis, M., Mansell, R. and Van den Bulck, H. (ed.) *Handbook of Media and Communication Governance* Edgar Elgar.

³⁷⁸ Michalis, M. "Competition Law and Regulation." Edgar Elgar, 2024, p. 7.

³⁷⁹ On the nature of this duty of care, see Lorna Woods, 'The Duty of Care in the Online Harms White Paper' (2019) 11 *Journal of Media Law* 6.

³⁸⁰ Kira, Beatriz, Pro-Competition and Online Safety: A More Holistic Approach to Digital Platform Regulation (January 7, 2024). Maria Ioannidou and Despoina Mantzari (eds.), *Research Handbook on Competition Law and Data Privacy*, Edward Elgar Publishing (2024) p.11.

comparison between the UK OSA and the EU DSA concerning their approaches to hate speech and content moderation. The OSA introduces a 'duty of care' model, shifting from intermediary liability to proactive harm prevention, and emphasises the need for platforms to protect users, particularly vulnerable groups, by conducting risk assessments and ensuring transparency. This contrasts with the EU's DSA, which includes the 'Good Samaritan' rule, offering liability protections for platforms that voluntarily remove harmful content and focuses on systemic risks and transparency. The paragraph highlights the ongoing debate about the responsibility of social media platforms, the evolution of regulations for content moderation, and the implications of these regulations for online platforms.

4.6 Conclusions

This chapter has aimed to address how the legal frameworks of UK manage the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression. Through a detailed comparative analysis of the UK's Online Safety Act (OSA) and the EU's Digital Services Act (DSA), we have examined the various regulatory approaches and their implications. A key concern emerges about whether the UK law will significantly influence content moderation practices beyond handling illegal content. Initially, the Online Safety Bill (OSB) included a category for 'legal, but harmful content' aimed at adults, encompassing potentially harmful yet legal material like misogynistic abuse. However, this category faced criticism for its vague definition, potential government intervention in legal speech, and concerns about impeding free speech. Consequently, the UK government removed this category, narrowing the OSA's scope. Instead, Category 1 platforms now must provide 'user empowerment' tools, which, though basic, align with existing features on major platforms. Questions persist about the effectiveness of these tools in preventing indirect harm and whether default settings adequately protect users. The OSA's focus on specific content types limits its authority in addressing systemic risks in content moderation systems compared to European law, which requires a broader assessment of risks, including those related to well-being, public health, political processes, and fundamental rights. Additionally, the OSA lacks a crisis response mechanism found in the DSA, restricting Ofcom's oversight during critical moments in content moderation. The

phrase 'online harm' is frequently used in current debates about online safety regulation, but it is often assumed to have a clear and agreed-upon meaning, overlooking the fact that there's considerable debate and disagreement about what exactly constitutes 'online harm'³⁸¹. Proposals for regulating online safety, especially those broadening the definition of 'harm' beyond traditional legal boundaries (such as including a category for 'legal but harmful' content), have raised concerns among various stakeholders about the risk of excessive regulation. In response to these concerns, this dissertation presents two main arguments. First, relying solely on narrow legal definitions of harm might result in overlooking important issues. This is because existing legal frameworks may not adequately address the cumulative harms experienced by historically marginalised groups and broader societal impacts, especially considering the unique dynamics of digital platforms and the emergence of new forms of harm facilitated by these platforms. Additionally, expanding the scope of online harm definitions in safety regulations could open avenues for more comprehensive solutions. While platforms currently address 'legal but harmful' content or behaviour, they often do so selectively and reactively, typically in response to highly publicised incidents³⁸², and with minimal transparency and public oversight³⁸³. Regulators have indicated a readiness to allow platforms to explore various solutions to tackle legal harm beyond simply removing content and banning users. These approaches are often influenced by existing criminal justice systems but aim to be more nuanced and flexible³⁸⁴. The effectiveness of these approaches will depend on their alignment with the specific types of harm they aim to mitigate. For instance, if content like subtle racism, such as the use of banana emojis targeting black individuals, is deemed harmful, relying solely on user-level controls may protect users from being directly abused but may not tackle the broader societal impact of such actions that

³⁸¹L Bartolo and A Matamoros Fernandez (n 282) 1-16.

³⁸²Ananny, Mike, and Tarleton Gillespie. "Public platforms: Beyond the cycle of shocks and exceptions." *IPP2016 The Platform Society*, 2016, pp.1-22.

³⁸³L Bartolo and A Matamoros Fernandez (n 282) 1-16.

³⁸⁴Woods, Lorna. "Online harm reduction—a statutory duty of care and regulator." *Available at SSRN 4003986*, 2019.

perpetuate racism. Government regulation promoting "procedural accountability"³⁸⁵ can ensure that platforms are responsible for managing lawful online harms effectively. This approach offers a middle ground between regulators overlooking harms that don't meet legal standards and unnecessarily turning to criminalisation. Requiring platforms to maintain a suitable level and type of transparency is essential to thoroughly assess the efficacy of various measures. Provisions related to "lawful but harmful" content in the OSB, not included in the OSA, which faced considerable criticism, did not mandate specific actions to address such content or behaviour. Instead, they required platforms to assess risks and transparently outline their approach to managing these risks in their terms of service. Consequently, the Bill granted platforms a level of flexibility that, theoretically, could prevent excessive regulation and content removal. It was drafted in a manner that didn't compel platforms to remove or limit "lawful but harmful" content or behaviour but mandated them to document their approach to dealing with such harm. The final version of the Online Safety Act, which was passed into law in October 2023, eliminated specific obligations for platforms to conduct risk assessments or take actions regarding "legal but harmful" content accessible to adults. Instead, the focus of the OSA shifted primarily to addressing illegal content and content harmful to children. Platforms are now required to provide users with tools to manage their exposure to certain types of lawful but potentially harmful content and to enforce their own terms of service. This means that while platforms still have some responsibility to deal with harmful content, the strict regulatory obligations initially proposed were not included in the final Act. Relying solely on strict legal definitions of harm is only a partial and ultimately constraining approach. Regulating online safety to address lawful harms and requiring platforms to adopt an organised, inclusive, and transparent approach offers a vital opportunity to address a wide range of issues and prioritise the views of historically marginalised communities in platform management.

³⁸⁵*Ibid.*

Chapter 5: Conclusion and Recommendations

5.1 Conclusion

This final chapter aims to draw conclusions from the preceding discussion and offer recommendations based on the findings presented in the dissertation. The Digital Services Act (DSA) represents a substantial and inclusive proposal designed to establish fundamental regulations and principles governing the intermediary role in digital content dissemination. As discussed in section 3.5, this legislative framework introduces significant user rights and service provider obligations, particularly focusing on Very Large Online Platforms (VLOPs). These obligations include enforcing terms and conditions, ensuring transparency, defining criteria for content removal, establishing grievance mechanisms, and implementing alternative dispute resolution mechanisms, among others. However, concerns arise regarding the duties and responsibilities associated with identifying and mitigating systemic risks, which may potentially impede users' freedom of expression. To address these concerns, particularly in the context of Articles 26 and 27 of the DSA, several considerations warrant attention. These include the constraining nature of the systemic risk identification/mitigation mechanism, the need to refine amendments targeting specific illicit content rather than broad categories of harmful content, the reinforcement of liability exemptions over the imposition of due diligence obligations, the necessity for reasonable and transparent measures to curb the dissemination of illegal content, and ultimately, the importance of conducting robust and comprehensive human rights impact assessments. Similarly, the enactment of the Online Safety Act (OSA) in 2023

represents a significant step toward holding online service providers accountable for combating illicit and harmful content. However, revisions are necessary to safeguard against potential infringements on freedom of expression and other foundational rights. The existing regulatory framework in the United Kingdom already imposes substantial constraints, as evidenced by notice-and-action regimes and government-mandated content moderation. Government-mandated content moderation under the OSA includes legal obligations for online platforms to protect users from illegal and harmful content, requirements for platforms to have systems to detect and remove illegal content using both automated tools and human moderation, regular assessments to identify potential risks and implement mitigation measures, the production of transparency reports detailing their content moderation efforts, provision of mechanisms for users to report illegal or harmful content, and regulatory oversight by Ofcom to ensure compliance, conduct audits, impose fines, and issue codes of practice. Additionally, platforms are required to implement transparency and auditing obligations rooted in algorithms for content moderation and platform recommendation systems to ensure these tools are not biased or leading to over-removal of lawful content. Thus, it is imperative to effectively harmonise online safety imperatives with the imperative of upholding freedom of expression. Drawing on the findings of this study, five distinct sets of revisions to the OSA have been proposed to address its most significant deficiencies. These include clarifying service providers' obligations regarding content moderation measures, establishing an effective and transparent redress system, devising a strategy to protect freedom of expression across all forms of content, introducing transparency and auditing obligations rooted in algorithms for content moderation and platform recommendation, and further refining the definition of illegal and harmful content. Moreover, statutory frameworks, such as the European Union's DSA and the United Kingdom's OSA include detailed notice-and-action procedures, which require platforms to remove illegal content promptly upon notification. Additionally, these frameworks impose state-mandated oversight mechanism to ensure compliance, such as the requirement for VLOP's to conduct annual risk assessments to identify and mitigate systemic risks associated with illegal content dissemination, as well as algorithmic transparency measures to prevent

biases and over-removal of lawful content. However, the prevailing regulatory landscape in the United Kingdom may be characterised more accurately by terms such as "restricted," "permitted," "proscribed," and "censored," given the considerable influence exerted by factors such as the identity of the speaker, the nature of the content, and the platform of expression. The OSA extends beyond conventional notice-and-action mechanisms by requiring businesses to actively seek out and curtail online content, potentially leading to algorithm-driven censorship. This approach raises concerns regarding accountability, transparency, and adherence to democratic principles. Furthermore, the study identifies challenges in online content moderation, particularly in delineating responsibility for formulating enforceable and efficacious regulatory frameworks. These challenges underscore the complexities inherent in regulating internet content and highlight key objectives in governing the internet and regulating online content, including fostering a free, open, and secure online environment, preventing internet abuse, implementing balanced oversight of commercial platform usage, and averting governmental exploitation of regulatory measures for surveillance, censorship, and dissent suppression. This research has addressed the central question: "How do the legal frameworks of the EU and UK address the liability of internet intermediaries in regulating online hate speech while balancing the right to freedom of expression?" The findings highlight the delicate balance required to manage hate speech effectively while ensuring that regulatory measures do not unduly infringe on freedom of expression.

Addressing these challenges entails tackling a global regulatory deficit, rectifying inconsistencies in moderation and enforcement, improving the detection and removal of harmful content, especially implicit or encoded content, and devising strategies to manage the enduring presence of digital content over time.

5.2 Recommendations

The governance of Digital Service Providers (DSPs) in contemporary information societies presents numerous ethical and societal challenges, as discussed in Chapter 2, due to their central role in

facilitating information accessibility and shaping interactive dynamics, thereby serving a fundamental democratic function. Prioritising fundamental rights in discussions regarding DSP regulation is crucial, as they serve as indispensable safeguards for democracy and the rule of law within EU and UK legal frameworks. This emphasis on fundamental rights aims to create an environment where users can express themselves freely while upholding human rights standards. Previous regulatory efforts have aimed to strike a balance between users' rights and DSP responsibilities. However, the novel Digital Services Act (DSA) seeks to strengthen the protection of fundamental rights while maintaining regulatory effectiveness. Yet, the complexity of the DSA and the preference for national oversight strategies over a unified European approach may pose challenges in achieving consistent implementation, necessitating subsequent revisions to align the DSA with the rapidly evolving digital landscape. The DSA seeks to maintain the equilibrium established by preceding legislation, such as the E-Commerce Directive (ECD), emphasising the importance of fundamental rights in the online sphere. Balancing fundamental rights should be guided by principles outlined in the European Convention on Human Rights (ECHR) and interpreted through jurisprudence from the European Court of Human Rights (ECtHR) and the Court of Justice of the European Union (CJEU). Various challenges must be addressed, including avoiding excessive restrictions on freedom of expression, protecting privacy and data, ensuring fairness in content moderation decisions, and preventing disproportionate obligations that may hinder DSP business freedoms. To safeguard fundamental rights, the DSA proposes establishing specialised oversight bodies at the European level to monitor DSP compliance. These bodies should collaborate closely with the European Union Agency for Fundamental Rights to ensure transparency and human oversight in content moderation. Additionally, the DSA aims to ensure that DSP obligations are proportionate to their technical and operational capacities, thus avoiding undue burdens on entrepreneurial freedoms. Collaborative efforts should involve consulting independent experts to reform platform policies and functionality, promoting healthy discourse and benefiting both platform owners and users. Enhancing transparency and user control is essential, as is safeguarding the right to freedom of expression for all individuals, regardless

of content origin. Governments must address communication challenges through nuanced policies rather than onerous regulations, employing strategies to combat hate speech while upholding freedom of expression. Regulations should aim to counter monopolistic control without censoring dissenting views, and specific measures are needed to regulate automatic content identification processes to prevent overly broad censorship risks. Governments should prioritise transparency, strengthen research efforts, and ensure users' access to meaningful redress mechanisms. Citizens also have a responsibility to critically evaluate information and engage ethically, particularly on social media platforms, to combat hate speech and foster healthy discourse. The UK must consider the implications of content blocking without due diligence on harmfulness, as such measures risk infringing upon the right to free speech. Insights from the European experience, as demonstrated by the DSA, can inform the UK's approach to harmonising online safety with freedom of expression. In summary, the DSA proposes a nuanced and proportionate approach to DSP regulation, aiming to protect fundamental rights online without stifling innovation and business freedom in the digital realm.

Bibliography

- A. Brouillette, 'Key Findings: Companies are Improving in Principle but Failing in Practice' 2020 Ranking Digital Rights Corporate Accountability Index
<https://rankingdigitalrights.org/index2020/key-findings>.
- Akbarova, Lala. "Hate Speech-Legal, Political, Cultural Aspects. World Experience In Legal Regulation." Distance Education in Ukraine: Innovative, Normative-Legal, Pedagogical Aspects 2: 22-44.
- Alexander, Sam. "A uniquely Australian approach: A thematic analysis of the normative foundations of Australia's approach to the regulation of the internet." *Adel. L. Rev.* 43 (2022): 345.
- Alkiviadou, Natalie. "The legal regulation of hate speech: The international and European frameworks." *Politická misao*, 2018.
- Anand, R. and Bird, B. (2023). 'The Digital Services Act and Vulnerability: How Vulnerability is Defined by Digital Regulations in the EU'. *Irish Law Times*, 41(9), pp. 135-140. Available at: <https://doi.org/10.34961/researchrepository-ul.24064260.v1> (Accessed: 23 January 2024)
- Angel N Desai, Diandra Ruidera, Julie M Steinbrink, Bruno Granwehr, Dong Heun Lee, Misinformation and Disinformation: The Potential Disadvantages of Social Media in Infectious Disease and How to Combat Them, *Clinical Infectious Diseases*, Volume 74, Issue Supplement_3, 15 May 2022. <https://doi.org/10.1093/cid/ciac109>.
- Angelopoulos and S. Smeth, 'Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability', (2017) 8(2) *Journal of Media Law*, 266–301 <https://www.ivir.nl/publicaties/download/Notice_and_Fair_Balance.pdf>.
- Ananny, Mike, and Tarleton Gillespie. "Public platforms: Beyond the cycle of shocks and exceptions." IPP2016 The Platform Society (2016).
- Baccarella, Christian V., Timm F. Wagner, Jan H. Kietzmann, and Ian P. McCarthy. "Social media? It's serious! Understanding the dark side of social media." *European Management Journal* 36, no. 4 (2018): 431-438.
- Bahl, Varun Sen, et al. "Internet intermediaries and online harms: Regulatory." (2020).
- Balkin, J. M. (2018). *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*. *UC Davis Law Review*, 51, pp.1149-1210.

- Barker, Kim, et al. "Draft Online Safety Bill. Written evidence submitted by the British and Irish Law, Education and Technology Association (BILETA)(OSB0073)." (2021).
- Bartolo, L., and A. Matamoros-Fernández. "Online harms. Platform governance terminologies essay series." Information Society Project, Yale Law School. <https://law.yale.edu/isp/publications/platform-governance-terminologies> (2023).
- Bassini, Marco. "Fundamental rights and private enforcement in the digital age." *European Law Journal* 25.2, 2019.
- Bello, Barbara Giovanna. "Tackling online hate speech from a European perspective: Potentials and challenges of inter-legality." *Oñati Socio-Legal Series*, 2023.
- Bello, B.G., 2021. I discorsi d'odio in rete. In: T. Casadei and S. Pietropaoli, eds., *Diritto e tecnologie informatiche: Questioni di informatica giuridica. Prospettive, istituzionali e sfide sociali*. Padua: Cedam, 247–262.
- Bello, B.G., and Scudieri, L., 2022. Discorsi d'odio online. Spunti per un dibattito interdisciplinare. In: B.G. Bello and L. Scudieri, eds., *L'odio online: forme, prevenzione e contrasto*. Turin: Giappichelli, 1–18.
- Bertolini, Andrea, Francesca Episcopo, and NICOLETA-ANGELA Cherciu. "Liability of online platforms." (2021): 1-149.
- Blumenthal, Richard, and Marsha Blackburn. "Kids Online Safety Act of 2023." Retrieved January 5 (2023): 2024.
- Boyd, Danah. "Social network sites as networked publics: Affordances, dynamics, and implications." *A networked self*. Routledge, 2010. 47-66 (Zizi Papacharissi ed., 2010).
- B. PETKOVA, T. OJANEN (Eds), *Fundamental Rights Protection Online, The Future Regulation of Intermediaries*, Cheltenham- Northampton, 2020.
- Brown, Nina I. "Regulatory Goldilocks: Finding the Just and Right Fit for Content Moderation on Social Platforms." *Tex. A&M L. Rev.* 8 (2020): 451.
- Bryce, India, et al. "When quantity takes on a quality of its own: A retrospective exploration of the lived experience of cumulative harm." *Child abuse review* 32.4 (2023): e2800.
- Buiten, Miriam C. "The Digital Services Act From Intermediary Liability to Platform Regulation." *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12 (2021): 361.

- Buri, Ilaria, and Joris van Hoboken. "The Digital Services Act (DSA) proposal: a critical overview." Digital Services Act (DSA) Observatory (2021).
- Citron, D. K., & Wittes, B. (2017). "The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity." *Fordham Law Review*, 86(2).
- Claussen, Victor. "Fighting hate speech and fake news. The Network Enforcement Act (NetzDG) in Germany in the context of European legislation." *Media Laws* 3.3 (2018): 110-136.
- Cliquennois, Gaëtan, Sonja Snacken, and Dirk van Zyl Smit. "Can European human rights instruments limit the power of the national state to punish? A tale of two Europes." *European Journal of Criminology*, 2021 pp. 11-32
- Cobbe, Jennifer, and Jatinder Singh. "Regulating recommending: motivations, considerations, and principles." *Considerations, and Principles* (April 15, 2019) (2019).
- Coe, Peter. "The Draft Online Safety Bill and the regulation of hate speech: have we opened Pandora's box?" *Journal of Media Law* 14.1 (2022), pp. 50-75.
- Cohn, Cindy. "Bad Facts Make Bad Law: How Platform Censorship Has Failed So Far and How to Ensure That the Response to Neo-Nazis Doesn't Make It Worse." *Geo. L. Tech. Rev.* 2 (2018): 432-439.
- Communications Decency Act (CDA), Pub. L. No. 104-104 (Tit. V), 110 Stat. 133 (Feb. 8, 1996).
- Council and Parliament Directive 2000/31/EC of 8th June 2000 (E-Commerce Directive) [2000] OJ L 178/1, Article 15(1)
- Council of Europe. (2003). Additional Protocol to the Convention on Cybercrime, concerning the criminalisation of acts of a racist and xenophobic nature committed through computer systems (ETS No. 189). Strasbourg.
- Council of Europe, Committee of Ministers Recommendation No. R (97) 20
- Council of Europe. (1950). *European Convention on Human Rights*. Article 10.
- Council of Europe Recommendation CM/Rec (2022)16 of the Committee of Ministers to member States on combating hate speech, para.2 (Adopted by the Committee of Ministers on 20 May 2022 at the 132nd Session of the Committee of Ministers).

- Council of Europe Recommendation on the Roles and Responsibilities of Internet Intermediaries (CM/Rec (2018)2) (Adopted by the Committee of Ministers on 7 March 2018 at the 1309th meeting of the Ministers' Deputies).
- Council Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law, 28 November 2008 (2008/913/JHA).
- Council and Parliament Directive 2000/31/EC of 8th June 2000 (E-Commerce Directive) [2000] OJ L 178/1 , Article 15(1).
- Craig, P., & de Búrca, G. (2020). *EU law: Text, cases, and materials* (7th ed.). Oxford University Press.
- Daphne Keller, 'Broad Consequences of a Systemic Duty of Care for Platforms', Stanford Law School, Centre for Internet and Society, 1 June 2020.
- Davis III, Charles HF. Dream defending, on-campus and beyond: A multi-sited ethnography of contemporary student organising, the social movement repertoire, and social movement organization in college. Diss. The University of Arizona, 2015.
- Deacon, Daniel T. "Institutional Considerations for the Regulation of Internet Service Providers." *Berkeley Tech. LJ* 37 (2022): 309.
- De Morais, Carlos Blanco. "'Digital Democracy': A Threat to the Democratic System or Oxygenation of Representative Democracy and Free Speech?." *The Rule of Law in Cyberspace*. Cham: Springer International Publishing, 2022. 9-63.
- Department for Digital, Culture, Media and Sport, Online Harms White Paper: "Full government response to the consultation" (CP 354, 15 December 2020).
- DE STREEL, Alexandre. "Online Platforms' Moderation of Illegal Content Online-Law, Practices." (2020).
- Dias, T., "Amending Online Safety Bill to Ensure Consistency With Core International Human Rights Instruments: Specific Recommendations." <https://www.elac.ox.ac.uk/wp-content/uploads/2022/03/hatespeechandtheonlinesafetybill-supplementaryevidencesubmission-diaspdf.pdf>
- Dias, Talita. "Tackling online hate speech through content moderation: The legal framework under the International Covenant on Civil and Political Rights." *Countering online hate and its offline consequences in conflict-fragile settings* (Forthcoming) (2022).

- 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 1-16.
- 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, recital 47 and Art. 25, Measures against websites containing or disseminating child pornography.
- Dittel, Alexander. "The UK's Online Safety Bill: The day we took a stand against serious online harms or the day we lost our freedoms to platforms and the state?." *Journal of Data Protection & Privacy* 5.2 (2022): 183-194.
- DMCA (The Digital Millennium Copyright Act), Pub. L. No. 105-304, 112 Stat. 2860 (Oct. 28, 1998);
- D. KEATS CITRON, 'What to Do about the Emerging Threat of Censorship Creep on the Internet' (2017), Cato Institute
- D. Keller, 'The DSA's Industrial Model for Content Moderation' (2022) *Verfassungsblog* <https://verfassungsblog.de/dsa-industrial-model/>
- D. Keller, 'The EU's new Digital Services Act and the Rest of the World' (*Verfassungsblog*, 7 November 2022) <<https://verfassungsblog.de/dsa-rest-of-world>>.
- Duivenvoorde, Bram, and Catalina Goanta. "The regulation of digital advertising under the DSA: A critical assessment." *Computer Law & Security Review* 51 (2023): 105870.
- Duranti, Marco. *The conservative human rights revolution: European identity, transnational politics, and the origins of the European Convention*. Oxford University Press, 2016.
- E. DOUEK, 'U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for Human Rights by Default' (2018).
- E. M. ASWAD, 'The Future of Freedom of Expression Online' (2018) 17 *Duke Law & Technology Review* 1, 52-53.
- European Commission. *Code of conduct on countering illegal hate speech online*. European Commission document. 2016, p.1-23. http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf. Accessed August 12, 2017.

- European Commission, 2021. Communication on Guidance on Strengthening the Code of Practice on Disinformation (2021) p.262 final) [online]. 26 May. Available from: <https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2021:262:FIN> [Access 2 November 2022].
- European Commission Factsheet, ‘Code of Conduct on Countering Illegal Hate Speech Online: First Results on Implementation’, at 1 (Dec. 2016), http://ec.europa.eu/information_society/newsroom/image/document/2016-50/factsheet-code-conduct-840573.pdf
- European Commission. Strengthened Code of Practice on Disinformation [online]. 16 June. Brussels, 2022 p. 14. Available from: <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation> [Access 2 January 2023].
- European Commission, “The Project CONTRO” 2021. Available at: <https://ec.europa.eu/newsroom/just/items/704254>; OHCHR, “I Don’t Forward Hate: an online campaign against hate speech in Pakistan.” 2019. <https://www.ohchr.org/en/get-involved/stories/i-dont-forward-hate-online-campaign-against-hate-speech-pakistan>
- E. M. Aswad, ‘The Future of Freedom of Expression Online’ (2018) 17 Duke Law & Technology Review 1, pp. 52-53.
- Eriksen, Erik Oddvar, and John Erik Fossum. "Democracy through strong publics in the European Union?" JCMS: Journal of Common Market Studies 40, no. 3 (2002): 401-424.
- European Parliament and Council Directive 2018/1808 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, 14 November 2018 (PE/33/2018/REV/1). Official Journal, L 303, 28.11.2018, 69–92.
- European Parliament and Council Directive 2018/1808 of 14th 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 (2018) OJ L 303/69 (from herein referred to as the ‘AMSD’), recital 45.
- European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC. Official Journal of the European Union.

- Evelyn Mary Aswad, 'The Future of Freedom of Expression Online' (2018) 17 Duke Law & Technology Review 1, pp. 52-53.
- F. Caloprisco, The Court of Justice expresses itself on the territorial scope of the obligation to de-index personal data online, in *Annali Aisdue*, vol. I, p. 357 ff.
- Feinberg, Joel. "The moral limits of the criminal law: volume 3: harm to self." Oxford University Press, USA, 1989.
- F. Erixon, 'Too Big to Care or Too Big to Share: The Digital Services Act and the Consequences of Reforming Intermediary Liability Rules' (2021) European Centre for International Political Economy.
- Filatova-Bilous, Nataliia. "Content moderation in times of war: testing state and self-regulation, contract and human rights law in search of optimal solutions." *International Journal of Law and Information Technology* 31.1 (2023): 46-74.
- Fiorinelli, G., 2021. Ordinamenti interconnessi. Il contributo dell'interlegalità alla regolazione della rete. In: E. Chiti, A. di Martino and G. Palombella, eds., *L'era dell'interlegalità*. Bologna: Il Mulino, pp. 405–440.
- Fire, Michael, Roy Goldschmidt, and Yuval Elovici. "Online social networks: threats and solutions." *IEEE Communications Surveys & Tutorials* 16, no. 4 (2014): 2019-2036.
- Floridi, Luciano. *The onlife manifesto: Being human in a hyperconnected era*. Springer nature, 2015.
- Friedlaender, Christina. "On microaggressions: Cumulative harm and individual responsibility." *Hypatia* 33.1 (2018): 5-21.
- Friedlaender, Christina. "On microaggressions: Cumulative harm and individual responsibility." *Hypatia* 33.1 (2018), pp. 5-21.
- Frosio, Giancarlo. "Algorithmic enforcement online." *Intellectual Property and Human Rights (4th edition, Kluwer Law Int'l 2020)* (2019): p.709-744.
- Frosio, Giancarlo and Geiger, Christophe, Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime. *European Law Journal* 2023, Vol 29, Issue 1-2, pp. 31-77., Available at SSRN: <https://ssrn.com/abstract=3747756> or <http://dx.doi.org/10.2139/ssrn.3747756>

- Frosio, Giancarlo, and Martin Husovec. "Accountability and responsibility of online intermediaries." *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press, 2020) (2019).
- Galván, Nicolás Quaid. "Adopting the Cumulative Harm Framework to Address Second-Generation Discrimination." *Colum. J. Race & L.* 11 (2021): 147.
- Garcia-Teodoro, Pedro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. "Anomaly-based network intrusion detection: Techniques, systems and challenges." *computers & security* 28, no. 1-2 (2009): 18-28.
- Geiger, G. Frosio and E. Izyumenko, 'Intermediary Liability and Fundamental Rights' in Frosio (ed), n. 1, 138.
- Gentile, Giulia. "Between Online and Offline Due Process: The Digital Services Act." *New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights* (Springer, Forthcoming) (2023).
- Gergely Gosztanyi, How the European Court of Human Rights Contributed to Understanding Liability Issues of Internet Service Providers, 58 *ANNALES U. SCI. BUDAPESTINENSIS ROLANDO EOTVOS NOMINATAE* 121 (2019).
- Giatra, Efi Eirini. "The Digital Services Act: Making Europe fit for the Digital Age.", 2023.
- Gijrath, Serge, et al. "Concise European data protection, e-commerce and IT law." (2018): 1-1058.
- Gillett, Rosalie. "'This is not a nice safe space': investigating women's safety work on Tinder." *Feminist Media Studies* 23.1 (2023): 199-215.
- Gjoka, Minas, Michael Sirivianos, Athina Markopoulou, and Xiaowei Yang. "Poking Facebook: characterization of osn applications." In *Proceedings of the first workshop on Online social networks*, pp. 31-36. 2008.
- Goldman, Eric. "Content moderation remedies." *Mich. Tech. L. Rev.* 28 (2021): 1.
- Gongane, Vaishali U., Mousami V. Munot, and Alwin D. Anuse. "Detection and moderation of detrimental content on social media platforms: Current status and future directions." *Social Network Analysis and Mining* 12.1 (2022), p. 129.
- Gorenc, Nina. "Hate speech or free speech: an ethical dilemma?." *International Review of Sociology* 32.3 (2022): 413-425.

- Gosztonyi, Gergely. "How the European Court of Human Rights Contributed to Understanding Liability Issues of Internet Service Providers." *Annales U. Sci. Budapestinensis Rolando Eotvos Nominatae* 58 (2019): 121.
- GOV.UK. (2023). "Online Safety Act 2023." Legislation.gov.uk. <https://www.legislation.gov.uk/ukpga/2023/50/enacted>.
- Grabenwarter, Christoph. "The European Convention on Human Rights: Inherent Constitutional Tendencies and the Role of the European Court of Human Rights." *ELTE LJ* (2014).
- G'sell, Florence, The Digital Services Act (DSA): A General Assessment (April 15, 2023). in Antje von Ungern-Sternberg (ed.), Content Regulation in the European Union – The Digital Services Act, TRIER STUDIES ON DIGITAL LAW, Volume 1, Verein für Recht und Digitalisierung e.V., Institute for Digital Law (IRDT), Trier April 2023, Available at SSRN: <https://ssrn.com/abstract=4403433> or <http://dx.doi.org/10.2139/ssrn.4403433>
- Heldt, Amélie P. "EU Digital Services Act: The white hope of intermediary regulation." *Digital Platform Regulation: Global Perspectives on Internet Governance*. Cham: Springer International Publishing, 2022, pp. 69-84.
- Hillyard, Paddy, and Steve Tombs. "From 'crime' to social harm?." *Crime, law and social change* 48 (2007): 9-25.
- Hoffmann, Anna Lauren, and Anne Jonas. "Recasting justice for Internet and online industry research ethics." *Internet Research Ethics for the Social Age: New Cases and Challenges*. M. Zimmer and K. Kinder-Kuranda (Eds.), np Bern, Switzerland: Peter Lang, Forthcoming (2016).
- Höning, Nina. "The European Directive on e-Commerce (2000/31/EC) and its Consequences on the Conflict of Laws." *Global Jurist Topics* 5.2 (2005 p.12).
- Hoofnagle, Chris Jay, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. "The European Union general data protection regulation: what it is and what it means." *Information & Communications Technology Law* 28.1 (2019): 65-98.
- H. SCHULTE-NÖLKE ET AL., The legal framework for e-commerce in the Internal Market- State of play, remaining obstacles to the free movement of digital services and ways to improve the current situation, 2020, PE 652.707.
- Hui Zhen Gan, 'Corporations: The Regulated or the Regulators - The Role of IT Companies in Tackling Online Hate Speech in the EU' (2017) 24 *Columbia Journal of European Law* 111;

Gan deals with each in-depth in pages 118-125;
http://www.eesc.europa.eu/resources/docs/auto_coregulation_en--2.pdf

Igwebuike, Ebuka Elias, and Lily Chimunya. "Legitimizing falsehood in social media: A discourse analysis of political fake news." *Discourse & Communication* 15, no. 1 (2021): 42-58.

Isabelle Rorive, Rdguler l'Internet? [Regulating the Internet?], Address at the Conference on the Council of Europe's Convention on Information and Legal Co-operation Concerning "Information Society Services," May 21, 2003, available at http://www.coe.int/T/E/Legal-Affairs/Legal-co-operation/Information-Society-Services/Riguler%201%27Internet%20-%20CoE%20-%2021%20%20Mai%202003%20_version%20icrite_.pdf.

Isabelle Rorive, Strategies to Tackle Racism and Xenophobia on the Internet-WhereAre We in Europe?, 7 INT'L J. COMM. L. & POL'Y 8 (2003), available at <http://www.ijclp.org/72003/pdf/rorive-ijclp-artikel.pdf>;

Isabelle Rorive, What Can Be Done against Cyber Hate - Freedom of Speech versus Hate Speech in the Council of Europe, 17 CARDOZO J. INT'L & COMP. L. 417 (2009).

James Banks, 'European Regulation of Cross-Border Hate Speech in Cyberspace: The Limits of Legislation' 19 *European Journal of Crime, Criminal Law and Criminal Justice* (2011)

Jane, Emma A. "Misogyny online: A short (and brutish) history." (2016): 1-152.

Jasper P. Sluijs, From Competition to Freedom of Expression: Introducing Article 10 ECHR in the European Network Neutrality Debate, *Human Rights Law Review*, Volume 12, Issue 3, September 2012.

J. B. NORDEMANN, The functioning of the Internal Market for Digital Services: responsibilities and duties of care of providers of Digital Services, 2020, PE 648.802.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. "Does transparency in moderation really matter? User behaviour after content removal explanations on reddit." *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019): 1-27.

Jiang, Jialun Aaron, et al. "Understanding international perceptions of the severity of harmful content online." *PloS one* 16.8 (2021): e0256762.
J. MCHANGAMA, 'The War on Free Speech – Censorship's Global Rise' (2022) *Foreign Affairs* <<https://www.foreignaffairs.com/articles/world/2022-02-09/war-free-speech-censorship>>.

- J. MCHANGAMA, N. ALKIVIADOU AND R. MENDIRATTA, 'Thoughts on the DSA: Challenges, Ideas and the Way Forward through International Human Rights Law' (Justitia, 2022), p.8 https://futurefreespeech.com/wp-content/uploads/2022/05/Report_thoughts-on-DSA.pdf.
- J. Nathan Matias, Eric Pennington & Zenobia T. Chan, Testing Concerns about Technology's Behavioural Impacts with N-of-One Trials (2022).
- Joan Barata 'Regulating content moderation in Europe beyond the AVMSD' (Media@LSE Blog, February 25th, 2020) <https://blogs.lse.ac.uk/medialse/2020/02/25/regulating-content-moderation-in-europe-beyond-the-avmsd/>
- Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' Plataforma por la Libertad de Infoacion.
- John Mariette. "Defamation, the right to freedom of expression and libel reform in the UK: A comparative analysis of legal regulation and jurisprudence in England and Wales, the United States of America, Germany, and the European Court of Human Rights." Middlesex University (2022).
- Johnson, Brett Gregory. "Speech, harm, and the duties of digital intermediaries: Conceptualising platform ethics." *Journal of Media Ethics* 32, no. 1 (2017): 16-27.
- Johnson, David R., and David Post. "Law and borders: The rise of law in cyberspace." *stanford law review* (1996): 1367-1402.
- Johnson, Mark. "Save Online Speech Coalition-Online Safety Bill Committee Stage amendment briefing [May 2022]." (2022).
- JOHN STUART MILL, ON LIBERTY (1859).
- John Woodhouse, 'Online Safety Bill: Progress of the Bill' (House of Commons Library 2023) Research Briefing Number 9579.
- Jones, Mariette W. "Regulation of Online Speech in the UK." *University of the Pacific Law Review* 55.2 (2024): 261.
- Karjo, Clara Herlina, and Andreas Ng. "Hate Speech Propaganda from and against Muslims in Facebook Posts." *International Journal of Cyber Criminology* 14, no. 2 (2020): 400-416.
- 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), <http://data.europa.eu/eli/reg/2022/2065/oj>.

Kaushal, Vishal, and Manasi Patwardhan. "Emerging trends in personality identification using online social networks - a literature survey." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, no. 2 (2018): 1-30.

Kettemann, Matthias C. *The normative order of the internet: A theory of rule and regulation online*. Oxford University Press, 2020.

Kira, Beatriz, and Laura Schertel. "A Primer on the UK Online Safety Act." *Verfassungsblog*, DOI 10 (2023).

Kira, Beatriz. "Pro-Competition and Online Safety: A More Holistic Approach to Digital Platform Regulation." *Research Handbook on Competition Law and Data Privacy*, Edward Elgar Publishing, (January 7, 2024).

Kölbel, Tobias, and Daniel Kunz. "Mechanisms of intermediary platforms." *arXiv preprint arXiv:2005.02111* (2020).

Kozyreva, Anastasia, Stephan Lewandowsky, and Ralph Hertwig. "Citizens versus the internet: Confronting digital challenges with cognitive tools." *Psychological Science in the Public Interest* 21, no. 3 (2020): 103-156.

La Rue, Frank. "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression." *Human Rights Council* 16 (2011).

Leiser, Mark, and Dr Edina Harbinja. "Content not available: Why The United Kingdom's Proposal For A "Package Of Platform Safety Measures" Will Harm Free Speech." *Technology and Regulation* 2019 (2020): 78-90.

Leiser, Mark. "Reimagining Digital Governance: The EU's Digital Service Act and the Fight Against Disinformation." 2023, pp. 11-42.

Letsiou Häusler, Anna, Nicolas Beckenkamp, and Livia Röthlisberger. *New Dimensions of an Old Dilemma: Hate Speech and Freedom of Expression*. No. 7. Universitätsverlag Potsdam, 2019.

Levchak, Charisse C., and Charisse C. Levchak. "Microaggressions, macroaggressions, and modern racism." *Microaggressions and modern racism: Endurance and evolution* (2018): 13-69.

Livingstone, Sivernte. *The Delfi AS v Estonia judgement explained*. *London School of Economics and Political Science Blog*. (2015, p.22) <http://eprints.lse.ac.uk/id/eprint/80766>

- Llansó, E. et al. Artificial Intelligence, Content Moderation, and Freedom of Expression, TWG: Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Netherlands. 2020, p4.
<https://policycommons.net/artifacts/8986433/untitled/9875488/> (accessed on 28 May 2024).
- L. Niombo, T. Evas, Digital services act - European added value assessment, EPRS PE 654.180;
- Lodder, Arno R. "Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the internal market." *EU Regulation of E-Commerce*. Edward Elgar Publishing, 2017, pp. 15-58.
- Loucaides, Loukis G. "Questions of fair trial under the European Convention on Human Rights." *Hum. Rts. L. Rev.* 3 (2003): 27.
- Lorna Woods, 'The Duty of Care in the Online Harms White Paper' (2019) 11 *Journal of Media Law*.
- Lucas Volman, L. L. "Is the cornerstone loose? Critical analysis of the functioning of the 'country of origin's principle in the Audiovisual Media Services Directive, taking into account the rapid changes in the audiovisual industry and the recent challenges brought by Brexit." 2018, p.12.
- MacAvaney, Sean, et al. "Hate speech detection: Challenges and solutions." *PloS one*, 2019, pp.23-25.
- Maria Ioannidou and Despoina Mantzari (eds.), *Research Handbook on Competition Law and Data Privacy*, Edward Elgar Publishing, 2024, p.11, Available at SSRN: <https://ssrn.com/abstract=4631357>
- Maroni, Marta, and Elda Brogi. "Freedom of expression and the rule of law: the debate in the context of online platform regulation", *Research handbook on EU media law and policy*. Edward Elgar Publishing, 2021, pp. 166-189.
- Matias, Nathan, Eric Pennington, and Zenobia Chan. "Testing Concerns about Technology's Behavioural Impacts with N-of-one Trials." *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022.
- Mayson, Sandra G. "Bias in, bias out." *Yale LJ* 128 (2018) pp.18-21.
- McConnell-Ginet, Sally. "Why defining is seldom 'just semantics': Marriage and marriage." *Drawing the Boundaries of Meaning*. John Benjamins, 2006. 217-240.
- McCrudden, Christopher. "Human dignity and judicial interpretation of human rights." *European Journal of international Law* 19, no. 4 (2008): 655-724.

- Mchangama, Jacob, and Natalie Alkiviadou. "Hate speech and the European Court of Human Rights: Whatever happened to the right to offend, shock or disturb?." *Human Rights Law Review* 2021, pp. 13.
- McGonagle, T. The Council of Europe and internet intermediaries: A case study of tentative posturing. In R. K. N. Jayakar, V. M. Pickard, & J. T. Roberts (Eds.), *Global communications: Theories, stakeholders, and trends*, MIT Press. 2019, pp.11-14
- Mena, Jesus. Investigative data mining for security and criminal detection. Butterworth-Heinemann, 2003.
- M. Husovec, Holey Cap! CJEU Drills (Yet) Another Hole in the E-Commerce Directive's Safe Harbors, (2017) 12 (2) *Journal of Intellectual Property Law & Practice*, 125. <https://doi.org/10.1093/20-inplpw3>
- M. Husovec, Injunctions against Intermediaries in the European Union: Accountable but Not Liable? (CUP 2017).
- Michalis, M. "Competition Law and Regulation." Edgar Elgar, 2024.
- Milano, Silvia, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges." *Ai & Society* 35 (2020): 957-967.
- Mir, Joan Barata, and Marco Bassini. "Freedom of expression in the internet: main trends of the case law of the European Court of Human Rights." *The Internet and Constitutional Law*. Routledge, 2016, pp. 71-93.
- M. K, 'Code of Practice on Disinformation Revision Extended into 2022' (2022) Euractiv <https://www.euractiv.com/section/digital/news/code-of-practice-on-disinformation-revision-extended-into-2022/>
- Monti, Matteo. "The EU Code of Practice on Disinformation and the Risk of the Privatisation of Censorship." *Democracy and Fake News*. Routledge, 2020, pp. 214-225.
- Moshikaro, Khomotso. "Unjust Contempt as the Gist of Defamation Law." *Constitutional Court Review* 12.1 (2022): 59-86.
- M. Peguera, The DMCA Safe Harbors and Their European Counterparts: A Comparative Analysis of Some Common Problems, (2009) 32 (4) *Columbia Journal of Law & the Arts*, 481-512.
- Nash, Victoria. "Revise and resubmit? Reviewing the 2019 online harms white paper." *Journal of Media Law* 11.1 (2019): 18-27.

- Nathalie A. Smuha, Beyond the Individual: Governing AI's Societal Harm, 10 INTERNET POL'Y REV. (Sept. 30, 2021), <https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm>.
- Natali Helberger, Sarah Eskens et al, 'Implications of AI-driven Tools in the Media for Freedom of Expression' (Background Paper to the 2020 Cyprus Ministerial Conference, Council of Europe, September 2019) p 3 <https://www.ivir.nl/publicaties/download/AI-and-Freedom-of-Expression.pdf>.
- Neudert, Lisa-Maria. "Regulatory capacity capture: The United Kingdom's online safety regime." Internet Policy Review 12.4 (2023).
- N. Iacob, F. Simonelli, How to Fully Reap the Benefits of the Internal Market for E-Commerce?, 2020, PE 648.801.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022), 35 pages. <https://doi.org/10.1145/3457607>
- O'Connor, Margaret. "Documentary analysis and policy." *Research methods in palliative care* (2007), pp. 229-245
- Online Safety Bill 2022-23, HL Bill [87] (Rev) (U.K.)
- Oreste Pollicino, Judicial protection of fundamental rights in the transition from the world of atoms to the word of bits: the case of freedom of speech, (2019) 25 (2) European Law Journal, 168. <https://doi.org/10.1111/eulj.12311>.
- Ortolani, Pietro. "The Digital Services Act, Content Moderation and Dispute Resolution." Content Moderation and Dispute Resolution (February 1, 2023) (2023).
- Ost, François, and Michel Van de Kerchove. De la pyramide au réseau? Pour une théorie dialectique du droit. Vol. 94. Publications Fac St Louis, 2002.
- Pálmadóttir, Jóna Aðalheiður, and Iuliana Kalenikova. "Hate speech an overview and recommendations for combating it." Icelandic Human Rights Centre (2018): 1-27.
- Park, Tae Jung, and Akshita Rohatgi. "Balancing the platform responsibility paradox: A case for amplification regulation to mitigate the spread of harmful but legal content online." Computer Law & Security Review 52 (2024): 105960.

- Peguera, M. The Platform Neutrality Conundrum and the Digital Services Act. *IIC* 53, 2022, pp. 681–684. <https://doi.org/10.1007/s40319-022-01205-7>
- Perry, Barbara. "Exploring the community impacts of hate crime." *The Routledge international handbook on hate crime*. Routledge, 2014. 47-58.
- Persak, Nina. *Criminalising harmful conduct: the harm principle, its limits and continental counterparts*. Springer Science & Business Media (2007).
- Pollicino, Oreste, and Bassini, Marco, 2014. The Law of the Internet between Globalisation and Localisation. In: M. Maduro, K. Tuori and S. Sankari, eds., *Transnational Law: Rethinking European Law and Legal Thinking*. Cambridge University Press, 346–380.
- Pollicino, O., & Bassini, M. (2014). Free speech, defamation and the limits to freedom of expression in the EU: a comparative analysis. In *Research Handbook on EU Internet Law* (pp. 508-542). Edward Elgar Publishing
- Post, David G. "Governing Cyberspace: Law." *Santa Clara Computer & High Tech. LJ* 24 (2007): 883.
- P. Przemysław Polański, Rethinking, The notion of hosting in the aftermath of Delfi: Shifting from liability to responsibility?, *Computer Law & Security Review*, Volume 34, Issue 4, 2018, Pages 870-880, ISSN 0267-3649, <https://doi.org/10.1016/j.clsr.2018.05.034>.
- Press Release, Nadine Dorries, Secretary of State, UK Dep't for Digit., Culture, Media & Sport, Statement Made on 7 July 2022 [Statement UIN HCWS194] (July 7, 2022), pp.4-5 <https://questions-statements.parliament.uk/written-statements/detail/2022-07-07/hcws194>.
- Press Release, UK Department for Digital, Culture, Media & Sport & the Rt Hon Michelle Donelan MP, "New Protections for Children and Free Speech Added to Internet Laws" (Nov. 28, 2022), <https://www.gov.uk/government/news/new-protections-for-children-and-free-speech-added-to-internet-laws>.
- P. Van Eecke, 'Online Service Providers and Liability: A Plea for a Balanced Approach', (2011) *Common Market L. Rev.*, 1455, 1479–80.
- Quintel, T., & Ullrich, C. (2018). Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, related initiatives and beyond. In B. Petkova & T. Ojanen (Eds.), *Fundamental Rights Protection Online: The Future Regulation of Intermediaries*. Edward Elgar Publishing.

Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that constitutes Incitement to Discrimination, Hostility or Violence (2002) para. 22 Articles 19 and 20(2) ICCPR, in HRC documents, SRFOE reports and the Rabat Action Plan (RPA)

Ranking Digital Rights. Corporate Accountability Index. 2020

Recommendation CM/Rec(2018)2 of the Committee of Ministers to member States on the “roles and responsibilities of internet intermediaries”, 7th March 2018.

Regulation 2022/350 amending Regulation 833/2014 concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine. 2022/350

Regulation 2022/2065/EU 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), 2022 OJ L 277/1, <http://data.europa.eu/eli/reg/2022/2065/oj>.

Reidenberg, Joel R. "Lex informatica: The formulation of information policy rules through technology." *Tex. L. Rev.* 76 (1997): 553.

Report of the OSCE-ODIHR Expert Meeting, ‘Incitement to Hatred vs Freedom of Expression: Challenges of Combatting Hate Crimes Motivated by Hate on the Internet’ 2, 2010 <http://www.osce.org/odihr/68750?download=true>

Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, A/67/357 (7 September 2012).

Report of the Special Rapporteur on minority issues, Rita Izsák (A/HRC/28/64), Human Rights Council, Twenty-eighth session.

Reviglio, Urbano, and Claudio Agosti. "Thinking outside the black-box: The case for “algorithmic sovereignty” in social media." *Social Media+ Society* 6, no. 2 (2020): 2056305120915613.

Reynders, D. "6th Evaluation of the Code of Conduct. Factsheet [online]. European Commission, Directorate-General for Justice and Consumers, 7 October." (2021).

Roberts, Paul. "Does Article 6 of the European Convention on Human Rights require reasoned verdicts in criminal trials?." *Human Rights Law Review* 11, no. 2 (2011): 213-235.

SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

- Sally Broughton Micova, "What Is the Harm in Size?, CTR. ON REGUL. EUR." (Oct. 19, 2021), <https://cerre.eu/publications/what-is-the-harm-in-size>.
- Sander, Barrie. "Freedom of expression in the age of online platforms: The promise and pitfalls of a human rights-based approach to content moderation." *Fordham Int'l LJ* 43 (2019), p.939.
- S. Bardarova, Comparison Between the European Court of Justice and European Court of Human Rights, SSRN Paper, (18 June 2013) <https://doi.org/10.2139/ssrn.2281215>, <https://ssrn.com/abstract=2281215>.
- S. Becker, "Framing the Future of Internet", Social Europe, 2022, p.3 <<https://socialeurope.eu/branding-the-future-of-the-internet>>.
- Schauer, Frederick. "Fear, risk and the first amendment: Unraveling the chilling effect." *BUL rev.* 58 (1978), pp. 685-732
- Schreiner, M., & Hess, T. (2013). On The Willingness To Pay For Privacy As A Freemium Model: First Empirical Evidence. *ECIS 2013 Research in Progress*.
- Schwemer, Sebastian Felix, Digital Services Act: A Reform of the e-Commerce Directive and Much More (October 10, 2022). prepared for A Savin, Research Handbook on EU Internet Law (2022), Available at SSRN: <https://ssrn.com/abstract=4213014> or <http://dx.doi.org/10.2139/ssrn.4213014>
- Sebastian Becker, "Framing the Future of Internet" (2022) Social Europe <<https://socialeurope.eu/branding-the-future-of-the-internet>>.
- Sewell, Charlie. "More serious harm than good? An empirical observation and analysis of the effects of the serious harm requirement in section 1 (1) of the Defamation Act 2013." *Journal of Media Law* 12.1 (2020): 47-77.
- Sergio Andrés Castaño-Pulgarín, Natalia Suárez-Betancur, Luz Magnolia Tilano Vega, Harvey Mauricio Herrera López, Internet, social media and online hate speech. Systematic review, Aggression and Violent Behaviour, Volume 58, 2021, 101608, ISSN 1359-1789, <https://doi.org/10.1016/j.avb.2021.101608> (<https://www.sciencedirect.com/science/article/abs/pii/S1359178921000628#ab001>)
- S.-E. SKAANING & S. KRISHNARAJAN, 'Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech (2021) Justitia.

- Smith, P. (2006). The politics of UK television policy: the making of Ofcom. *Media, Culture & Society*, 28(6), 929-940. <https://doi.org/10.1177/0163443706068926>
- Smith, Steven D. "Is the harm principle illiberal." *Am. J. Juris.* 51 (2006): 1.
- Smith, Steven Douglas. "The Hollowness of the Harm Principle." Available at SSRN 591327 (2004).
- S.Morano-Foadi and L. Vickers - A matter for two courts: the fundamental rights question for the EU, Hart Publishing, 2015. 'Reflections on the Architecture of the EU after the Treaty of Lisbon: the European judicial approach to fundamental rights' in Special Issue edited by S Morano-Foadi and F Duina, entitled "'The Institutionalisation of Regional Trade Agreements Worldwide: New Dynamics and Future Perspectives', *European Law Journal*, Vol. 17, No. 5, September 2011, 595–610.
- S. Morano-Foadi and S. Andreadakis 'Protection of Fundamental Rights in Europe: The Challenge of Integration', Springer, June 2020.
- Smuha, Nathalie A. "Beyond the individual: governing AI's societal harm." *Internet Policy Review* 10.3 (2021).
- Söderlund, K., Engström, E., Haresamudram, K., Larsson, S., & Strimling, P. *Regulating high-reach AI: On transparency directions in the Digital Services Act*. *Internet Policy Review*, 13(1), 2024, p. 13, <https://doi.org/10.14763/2024.1.1746>.
- Șomlea, Radu, and Dr Caterina Sganga. "REGULATING INTERMEDIARY SERVICE PROVIDER." 2017, pp.1-29.
- Șomlea, Radu. "A Comparative Analysis of the EU and US Trends in Regulating Intermediary Service Provider Liability." *LL.M Short Thesis*, Central European University, 2017.
- Strand, Palma Joy, and Claudia E. Cohen. "Microaggressions and macro-injustices: How everyday interactions reinforce and perpetuate social systems of dominance and oppression." *Understanding and Dismantling Privilege* 12.2 (2022): 1-30.
- Štarienė, Lijana. "The limits of the use of undercover agents and the right to a fair trial under article 6 (1) of the European convention on human rights." *Jurisprudencija: mokslo darbai= Jurisprudence: Research Papers/Mykolas Romeris universitetas*. Vilnius: Mykolas Romeris universiteto Leidybos centras, 2009, Nr. 3 (117). (2009).
- Stuckey, Kent D. *Internet and online law*. Law Journal Press, 2023.

- Susan Benesch, 'But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies' (2020) 38 Yale Journal on Regulation Online Bulletin, p.90.
- Ten, Chin L., ed. *Mill's on liberty: A critical guide*. Cambridge University Press, 2008, p.2-6.
- Tarlach McGonagle 'The Council of Europe and Internet Intermediaries: A Case Study of Tentative Posturing' in Rikke Frank Jorgensen (ed), *Human Rights in the Age of Platforms* (MIT Press 2019), p. 235
- T. Dias Oliva, D. M. Antonialli & A. Gomes, 'Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online' *Sexuality & Culture*, 702; Facebook's Hate Speech Policies Censor Marginalised Users' *Wired*. 2021, 25
- T. Madiega, Reform of the EU liability regime for online intermediaries. Background on the forthcoming Digital Services Act, 2020, PE 649.404.
- Teleki, Cristina. *Due process and fair trial in EU competition law: The impact of article 6 of the European convention on human rights*. Brill, 2021.
- Ten, Chin L., ed. *Mill's on liberty: A critical guide*. Cambridge University Press, 2008.
- The Code of Conduct on Countering Illegal Hate Speech Online' (EU Commission, May 2016)
- Tosza, Stanisław. "Internet service providers as law enforcers and adjudicators. A public role of private actors." *Computer law & security review* 43 (2021): 105614.
- Udupa, Sahana. "Digital technology and extreme speech: Approaches to counter online hate." (2021).
- Unctad, 'Data Protection Regulations and International Data Flows: Implications for Trade and Development' (2016).
- UNOGPR Genocide Prevention and Responsibility 2024 pp. 1
<https://www.un.org/en/genocideprevention/> (accessed on 28 May 2024).
- United Nations (General Assembly).1948 "Universal Declaration of Human Rights (ICCPR)", December 1948.
- United Nations (General Assembly). 1966. "International Covenant on Civil and Political Rights." Treaty Series 999 (December): 171.

- United Nations Office on Genocide Prevention and the Responsibility to Protect (UNOGPR). Genocide Prevention and Responsibility, 2024. Available online: <https://www.un.org/en/genocideprevention/>.
- Van den Hoven, Jeroen, Martijn Blaauw, Wolter Pieters, and Martijn Warnier. "Privacy and information technology." (2014).
- Voorhoof, Dirk. "Freedom of expression and information and the case law of the European Court of Human Rights and the Court of Justice of the EU: overview and highlights 2014." *Justice for Free Expression in 2014, A review of global freedom of expression jurisprudence in 2014*. Columbia University, 2015.
- Voorhoof, D., "The Right to Freedom of Expression and Information under the European Human Rights System: Towards a More Transparent Democratic Society," *RSCAS 2014/12* (European University Institute, 2014).
- Wagner, Ben, et al. "Regulating transparency? Facebook, twitter and the German network enforcement act." Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020.
- Waldron, J., 2012. *The Harm in Hate Speech*. Cambridge, MA: Harvard University Press.
- Wanzo, Rebecca, and Reem Hilu. "Editors' Introduction: The Long History of Social Media." *Feminist Media Histories* 10.1 (2024): 1-16.
- Watney, Murdoch. "Regulation of social media intermediary liability for illegal and harmful content." European Conference on Social Media. Vol. 9. No. 1. 2022. GOV.UK. (2023). "Online Safety Act 2023." Legislation.gov.uk. <https://www.legislation.gov.uk/ukpga/2023/50/enacted>
- Watson, Gabrielle, Irit Tamir, and Brianna Kemp. "Human rights impact assessment in practice: Oxfam's application of a community-based approach." *Impact Assessment and Project Appraisal* 31, no. 2 (2013): 118-127.
- Weedon, Jen, William Nuland, and Alex Stamos. "Information operations and Facebook." Retrieved from Facebook: <https://www.mm.dk/wp-content/uploads/2017/05/facebook-and-information-operations-v1.pdf> (2017).
- Wendel, W. Bradley. "Explanation in legal scholarship: The inferential structure of doctrinal legal analysis." *Cornell L. Rev.* 96 (2010), p.1035

- Wendehorst, Platform Intermediary Services and Duties under the E-Commerce Directive and the Consumer Rights Directive, (2016) 5 (1) *Journal of European Consumer and Market Law*, 30-33.
- Willard, Nancy E. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research press, 2007.
- Wilman, Folkert. "The EU's System of Knowledge-Based Liability for Hosting Services Providers in Respect of Illegal User Content-Between the e-Commerce Directive and the Digital Services Act." *J. Intell. Prop. Info. Tech. & Elec. Com. L.* 12 (2021): 317.
- Wilson, Richard Ashby, and Molly K. Land. "Hate speech on social media: Content moderation in context." *Conn. L. Rev.* 52 (2020): 1029.
- Woods, Lorna. "Online harm reduction—a statutory duty of care and regulator." Available at SSRN 4003986 (2019).
- Woods, Lorna. "Regulating to Minimise Harm to Children and Young People." *Children, Young People and Online Harms: Conceptualisations, Experiences and Responses*. Cham: Springer International Publishing, 2024. Pp. 141-168.
- Wu, F.T., 2013. Collateral Censorship and the Limits of Intermediary Immunity. *Notre Dame Law Review*, 87(1), 293–349.
- Young, Jessica, Preetha Swammy, and David Danks. "Beyond AI: Responses to hate speech and disinformation." Unpublished paper presented at Carnegie Mellon University). Retrieved from <http://jessica-young.com/research/Beyond-AI-Responses-to-Hate-Speech-and-Disinformation.pdf> (2018).
- Ziccardi, G., 2019. Le espressioni d'odio online. In: G. Ziccardi and P. Perri, eds., *Tecnologia e Diritto*. Milan: Giuffrè Francis Lefebvre, 153–177.
- Ziccardi, Giovanni. *L'odio online: Violenza verbale e ossessioni in rete*. Vol. 102. Raffaello Cortina, Milano 2016.
- Ziccardi, G., and P. Perri. "L'odio online tra profilazione, big data e protezione dei dati personali." *L'odio online: forme, prevenzione e contrasto*. Giappichelli, 2022. 91-106.

Case Laws

Ahmet Yıldırım v Turkey App No 3111/10, (ECtHR, 18th December 2012)

C-682/18 and C-683/18, Frank Peterson v. Google LLC, YouTube Inc., YouTube LLC, Google Germany GmbH (C-682/18)

C-18/18 Glawischnig-Piesczek v Facebook Ireland Limited [2019] ECLI:EU:C:2019:821

C-236/08 Google France v Louis Vuitton Malletier SA [2008] ECLI:EU:C:2010:159 para 120

C-275/06 Promusicae v Telefónica de España SAU [2008] ECR I-271 para 70

C-291/13 Sotiris Papasavvas v O Fileleftheros Dimosia Etairia Ltd [2014] ECLI:EU:C:2014:2209

C-324/09 L'Oréal SA v eBay International AG [2011] ECR I-474.

C-557/07 LSG v Tele2 [2009] ECR I- 1227 para 29.

Delfi AS v Estonia App. no. 64569/09 (ECtHR, 16 June 2015) Para. 156.

Elsevier Inc. v. Cyando AG (C-683/18), EU:C:2021:503.

Erbakan v Turkey, App. no. 59405/00, (ECtHR, 6 July 2006). Para. 56.

Garaudy v France (dec.), App. no. 65831/01, (ECtHR, 24 June 2003).

Glimmerveen and Hagenbeek v. Netherlands, 4 E.H.R.R. 260, 18 D.R. 187 1982, pp. 3.

Gonzalez v. Google LLC, 598 U.S. 617 (2023).

Handyside v. United Kingdom (1976) (Application no. 5493/72).

Høiness v Norway App No 43624/14 (ECtHR, 9th March 2019).

Judgments of the Court of Justice of 2 December 2010, Case C-108/09, Ker-Optika bt, ECLI:EU:C:2010:725.

Judgments of the Court of Justice of 12 July 2011, Case C-324/09, L'Oréal SA, ECLI:EU:C:2011:474;

Judgments of the Court of Justice of 24 November 2011, Case C-70/10, Scarlet Extended, ECLI:EU:C:2011:771;

Judgments of the Court of Justice of 16 February 2012, Case C-360/10, SABAM v. Netlog NV, ECLI:EU:C:2012:85;

Judgments of the Court of Justice of 27 March 2014, Case C-314/12, UPC Telekabel Wien GmbH v. Constantin Film Verleih, EU:C:2014:192;

Judgment of the Court of Justice of 13 May 2014, Case C-131/12, Google Spain, ECLI:EU:C:2014:317.

Judgments of the Court of Justice of 15 September 2016, Case C-484/14, Tobias Mc Fadden, ECLI:EU:C:2016:689.

Judgments of the Court of Justice Cases C-236/08 to C-238/08 Google France SARL and Google Inc. v. Google France SARL and Google Inc, ECLI:EU:C:2010:159;

Judgment of the Court of Justice of 24 September 2019, Case C-507/17, Google LLC v. CNIL, ECLI:EU:C:2019:772.

Judgment of the Court of Justice of 3 October 2019, Case C-18/18, Eva Glawischnig- Piesczek v. Facebook Ireland, ECLI:EU:C:2019.

Lehideux and Isorni v. France, 2000 E.H.R.R.30 665 2000 pp. 1-12.

Magyar Tartalomszolgáltatók Egyesülete (MTE) and Index.hu Zrt v Hungary App No 22947/13 (ECtHR, 2nd February 2016).

Perinçek v Switzerland App no 27510/08 (ECtHR, 15 October 2015). Para. 114.

Pihl v Sweden App No 74742/14 (ECtHR, 9th March 2017);

Tamiz v the United Kingdom App No 3877/14 (ECtHR, 9th September 2017).

Twitter, Inc. v. Taamneh, 598 U.S. 471 (2023).