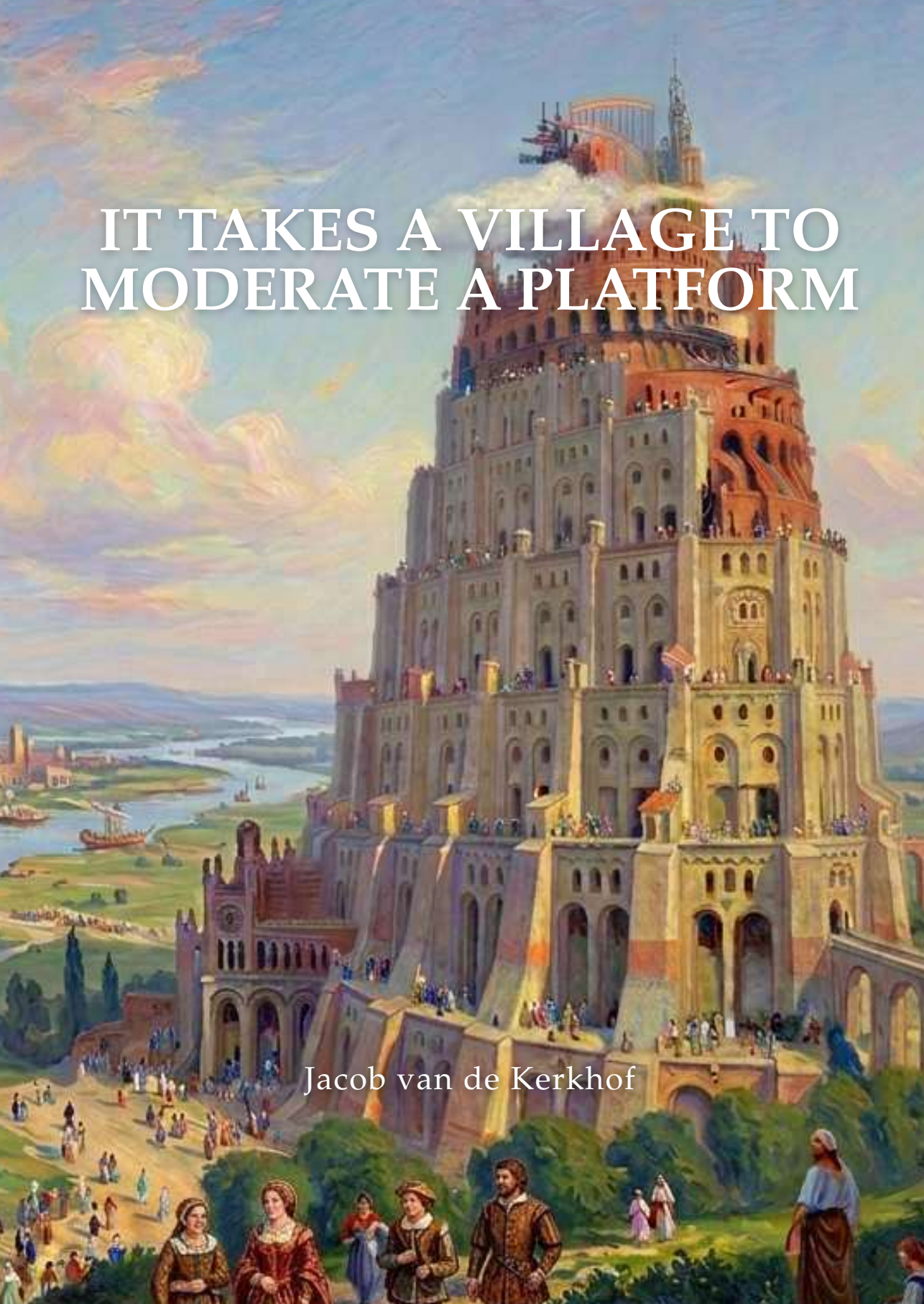
IT TAKES A VILLAGE TO MODERATE A PLATFORM



Jacob van de Kerkhof

It Takes a Village to Moderate a Platform

An Exploration of How Networked Platform Governance in the Digital Services Act affects Freedom of Expression

Jacob van de Kerkhof

It Takes a Village to Moderate a Platform

An Exploration of How Networked Platform Governance in the Digital Services Act affects Freedom of Expression

Vele handen maken licht werk: vrijheid van meningsuiting in gezamenlijk platformbestuur in de digitaledienstenverordening

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof. dr. ir. W. Hazeleger,
ingevolge het besluit van het College voor Promoties
in het openbaar te verdedigen op

vrijdag 29 mei 2026 des middags te 2.15 uur

door

Jacobus Johannes Wilhelmus van de Kerkhof
geboren op 12 oktober 1995
te Millingen aan de Rijn

Promotoren:

Prof. mr. dr. P.H. Blok
Prof. mr. dr. J.H. Gerards
Dr. E.C. Goanta

Beoordelingscommissie:

Prof. dr. J.F.T.M. van Dijck
Prof. dr. N. Helberger
Prof. dr. O. Pollicino
Prof. dr. S.F. Schwemer
Prof. mr. dr. S.A. de Vries

Table of Contents

Table of Contents 1
Acknowledgements 6
Chapter I: Introduction 11
1. Do Complex Questions Require Complex Answers? 11
2. Background 16
2.1. From e-Commerce Directive to Digital Services Act 16
2.2. Freedom of Expression in the Digital Realm 26
2.3. Multistakeholderism and Platform Governance 33
3. Approach, Outline, and Methodology 41
3.1. The Article-based Dissertation 41
3.2. Outline 41
3.2. Methodology 44
3.3. Contribution 46
3.4. Limitations 46
Part I: The Platform 51
Chapter II: Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database 53
1. Introduction 54
2. Prior work and contributions of the present study 55
3. Dataset 56
4. Statements of Reasons (SoRs) and DSA compliance 57
4.1. SoRs requirements 58
4.2. SoRs template and its compliance with the DSA 59
5. Platform practices and early DSA compliance 60
5.1. Who is Reporting: VLOPs 60
5.2. How Moderation Decisions and Detection are Done 62
5.3. Why: Reasons for Moderation 63
5.4. How Platforms Sanction: What Moderation Actions are Reported 66
5.5. Where Content is Moderated: Language and Territorial Scope 67
5.6. When SoRs are Submited (“Without Undue Delay”) 69
6. Conclusion: Does the DSA’s automated transparency rise to its promise? .72

Appendices 75
a. SoR schema design 75
Chapter III: To Host or not to Host: Rethinking Intermediary Liability Exemption in the Digital Services Act 81
Introduction 82
Section I: Intermediary Liability Exemption in the DSA 84
I.i An overview of conditional intermediary liability exemption 85
I.ii Why is the consensus that the hosting liability exemption has not changed? 94
I.iii: Intermediary findings 97
Section II: Why should we rethink the intermediary liability exemption of article 6 DSA? 97
II.i Insufficient access to functioning notice-and-action mechanisms 98
II.ii Fairer economic burden for hosting illegal content 100
II.iii Improved moderation capacity 101
II.iv Providing a countermovement to a trend of moderation withdrawal 102
II.v Freedom of expression is no *prima facie* obstacle to a more narrow interpretation 103
Section III: How does the DSA allow for rethinking the intermediary liability exemption? 105
III.i. An undiligent intermediary can be excluded from qualifying as a neutral hosting provider under article 6(1) 105
III.ii The threshold for awareness in diligent economic operators is lower 112
III.iii Synthesis 120
Conclusion 122
Part II: The User 123
Chapter IV: The Good, the Bad, the Compliant: Notice and Action Mechanisms under Article 16 DSA 125
1. Introduction 126
2. Background 127
2.1 What are notice-and-action mechanisms? 127
2.2 Article 16 DSA 129
2.3 Article 6 DSA 137
2.4 Summary 138

3. Methodology 139
3.1 The Walkthrough Method 139
3.2 Case selection and environment of expected use 140
3.3 Approach 142
4. Walkthrough and Article 16 Compliance 143
4.1 'Any individual or entity' 144
4.2 'Easily accessible' 146
4.3 User-friendliness 153
4.4 Facilitates sufficiently precise and adequately substantiated notices 156
4.5 Article 16 compliance 158
5. Implications of Article 16 Compliance for Platform Liability 163
6. Conclusion 165
Chapter V: Article 22 DSA: Building Trust with Trusted Flaggers 169
Introduction 170
Section I: The existing trusted flagger framework and its weaknesses 172
(i) Risk of over-removal 172
(ii) Fundamental rights and rule of law concerns 175
(iii) Transparency 178
Section II: Article 22 DSA 179
(i) Risk of over-removal 181
(i) Fundamental rights and rule of law concerns 183
(ii) Transparency 186
Section III: How can we better safeguard freedom of expression in – and ensure effective operationalisation of article 22 DSA? 188
(i) Harmonising selection criteria and procedure 188
(ii) Funding 190
(iii) Creating a harmonised API for trusted flaggers 190
Conclusion 192
Part III: The (Supra)State 195
Chapter VI: The DSA's Tower of Babel: on the Digital Services Coordinator and Freedom of Expression 197
I. Introduction 198

II: The Digital Services Coordinator 199
1. DSCs as an institution and their competence 199
2. Freedom of Expression 203
3. DSCs interacting with freedom of expression 206
4. Interim Conclusion 215
III: Decentralised enforcement of the DSA leading to different applications of the DSA 216
1. The DSA's decentralised enforcement structure 216
2. Different scope of freedom of expression facilitated by the DSA 218
3. Different procedures on a national level 220
4. Different enforcement strategies leading to different applications of the DSA 222
5. Interim Conclusion 225
IV. Suggestions for better safeguarding freedom of expression in DSA enforcement 226
1. Further centralise DSA enforcement 226
2. Further guide DSA enforcement 228
5. Conclusion 229
Chapter VII: Jawboning Content Moderation from a European Perspective 231
1. Introduction 232
2. Definitions: What Do We Mean by 'Jawboning'? 236
3. Freedom of Expression, Internet Intermediaries and the Rule of Law 239
3.1 Freedom of Expression and the European Court of Human Rights 239
3.2 Freedom of Expression and the European Union 242
3.3 Rule of Law as an Overarching Principle 244
3.4 Intermediate Conclusion 245
4. Jawboning, Content Moderation and the Rule of Law 246
4.1 Incentives for Jawboning in Content Moderation 246
4.2 What Does Jawboning in Content Moderation Look Like? 247
4.3 Constitutional Concerns Accompanying Jawboning in European Speech Governance? 249
5. Code of Conduct on Countering Illegal Hate Speech Online 250

Conclusion 254
Afterthoughts 256
Chapter VIII: Conclusion 259
1. The Internet Intermediary in Networked Platform Governance 260
2. Users and Civil Society in networked platform governance 263
3. The State 266
Synthesis: How does networked platform governance in the DSA affect the balancing of freedom of expression in content moderation? 269
Toward accountability in the DSA's networked platform governance 272
Discussion and final remarks 276
Bibliography 279
Regulation 279
Case law 281
Council of Europe case law 281
European Union case law 283
National case law 285
Literature 286
Varia 319
Vele handen maken licht werk: vrijheid van meningsuiting in gezamenlijk platformbestuur in de Digitaledienstenverordening 321
Curriculum Vitae 329
PhD Portfolio 330
PhD Articles 330
Other publications 330
Web-based output 331
Policy Input 332
Invited Participation 333

Acknowledgements

This dissertation is titled 'it takes a village to moderate a platform'. The same can be said about the very writing of this dissertation. I feel that the Dutch translation of the title, 'vele handen maken licht werk', more adequately portrays how I have felt over the course of this project. I would not have been able to get anywhere close to finishing this work without the many people helping, providing support, providing useful distraction, or – a personal favourite – disagreeing. This blessing makes this chapter simultaneously the easiest and most difficult chapter to write.

I am immeasurably indebted to my PhD supervisory team. I have often referred to you as my 'all-star team'. I have been incredibly fortunate to have been surrounded by some of the sharpest legal minds in the Netherlands. Peter, thanks for having an open mind to my misguided adventure of pursuing a PhD next to my Olympic ambitions and providing support at every turn to make the combination possible. Janneke, thank you for always taking the time to provide invaluable, incredibly detailed feedback at every turn. My work is only worth reading because of all the input you have provided. Stefan, thank you for believing in this research proposal, and helping me place the dissertation in the starting blocks. Catalina, I cannot express how grateful I am to have had you as a daily supervisor. It is rare to find people that are so committed, so engaged, and so creative in their field. It is even rarer to find people who are willing to share those capabilities, and you have been one of them.

To the HumanAds team, Taylor, Thales, Haoyang, Laura, Margje, Thijs: I am grateful to have been surrounded by such creative minds to write with, and vent to. Thanks for being great colleagues and great researchers. To the colleagues that are part of the Governing the Digital Society network: thank you for forming a place for rich multidisciplinary debate, and thank you for providing feedback on my work.

To all the people I have been fortunate enough to co-author with, Aleksandra, Anda, Giovanni, Jean, Jerry, Marlène, Rebeccah, Rishabh, Robert, Savvas: thank you for sharing your knowledge. Thank you for enriching my writing and thinking.

I have had the great fortune of being supported by amazing people that I have the privilege to call friends. Robbe, Joost, Piet, Paul, Jeroen, and Casper; thank you for tolerating me as your housemate for three years at our beloved Roeierspaleis de Houtman. We have dubbed ourselves the golden generation. No words cover our bond better.

As rowing has been a red thread throughout my life and throughout this PhD, it is only right that I thank the people that have been my teammates, for being more talented, stronger, more fun, and understanding towards my anger fits as a rower than I could ever be. Anish, Eise, Brecht, Casper, Max, Tijmen, Linda, Guido, Diederik, Mats, Martijn, Bas, Coen, Thomas, Harold, Kaj, Jakub, Fiep, Ralf, Haye, Rik, Michiel, Jan, Sander, Michiel, Nelson, Bram, Simon, Maarten, Dieuwke, Niki, Lennart, Abe, Mick, Guus, Guillaume, Ruben, Olav, Gert-Jan, Melvin, Pieter, Eli, Laila, Veronique: thank you. The same goes for the support that has been part of that journey, being coaches, trainers, physios, masseurs, psychologists, chefs, and others, in particular Ries, Coen, Niels, Thijs, Diederik, and Matthias.

In general, I owe a debt of gratitude to U.S.R. Triton and TeamNLRoeien as institutions that have set me up for an academic life aside from rowing. Only with your support and flexibility has this project been made possible.

I could not have been anywhere close to a doctorate without my parents supporting me at every step of the way. Thank you, pap en mam. I genuinely feel that I could not have had a better upbringing. That is, until I saw Eva, Jaap, Koos and Wies. Thank you for being part of my life.

No one has had to endure more of my stress, doubts, absence, business, bad jokes, and tiredness, and no one has endured them better, than my fiancée, Sabine. Thank you for being you. Thank you for having been part of this. Thank you for becoming my wife. Thank you for being my peace. Thank you.

"Tussen liefde aan de linkerkant

En rechts de eeuwigheid

En in ieder land een vaderland

De overkant

En jij"

'Trein, vuur, dageraad', Spinvis (2017)

Chapter I: Introduction

1. Do Complex Questions Require Complex Answers?

Mark Zuckerberg announced on 7 January 2025 to combat over-censorship and protect free speech on Meta platforms: "We built a lot of complex systems to moderate content. The problem with complex systems is, they make mistakes."[1] He felt that Meta's "good faith" attempts to combat illegal content online were stifling freedom of expression. In a political climate that, according to him, prioritises free speech more than ever, Zuckerberg felt that it was necessary to abandon some established content moderation practices in favour of free speech. For example, filtering algorithms would require higher confidence levels before removing content, only seriously illegal content would be removed without a user requesting it to be removed, and notoriously, Meta would rid itself of fact-checkers. Zuckerberg announced: "the fact-checkers have been too politically biased, and have destroyed more trust than they've created."[2] Independent fact-checkers are withdrawn in favour of a community notes system, in which users can add context to content in a democratised fashion. This is a trend across the industry. At the same time, Google indicated it would withdraw from fact-checking for YouTube, Google Ads, and its Search Engine, claiming that it already has "reasonable, proportionate, and effective measures to mitigate systemic disinformation risks on [its] services."[3] Big Tech's 'moderation withdrawal' follows a trend arguably started by Elon Musk's acquisition of X (then Twitter). After his acquisition he, amongst other far-reaching measures, dissolved the Twitter Trust and Safety Council, a group composed of civil society groups and academics that provided advise on how trust and safety could be maintained in Twitter.[4] The decisions by CEOs of American internet companies illustrate how the power of internet owners can affect how speech is governed on the internet, far beyond the borders of the United States.

These developments are crucial in shaping the online public discourse. Regulation across jurisdictions strikes varying balances in distributing power over speech on the internet, depending on national traditions on freedom of expression, economic

[1] Joel Kaplan, 'More Speech Fewer Mistakes' (*Meta,* 7 January 2025) <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>

[2] Ibid.

[3] Anupriya Datta, 'US Big Tech Backtracks on EU Fact-Checking Commitments' (*Euractiv*, 22 January 2025) <https://www.euractiv.com/section/tech/news/us-big-tech-backtracks-on-eu-fact-checking-commitments/> accessed 18 July 2025.

[4] Sheila Dang, 'Twitter Dissolves Trust and Safety Council' (*Reuters,* 13 December 2022) <https://www.reuters.com/technology/twitter-dissolves-trust-safety-council-2022-12-13/> accessed 18 July 2025.

interests, market power, and societal circumstances. In the European Union, that balance is struck in the Digital Services Act (Regulation 2022/2065, 'DSA').[5] The DSA creates a procedural framework through which speech online is governed. This procedural framework is aimed at resolving the struggle between state actors, citizens, and internet intermediaries on who decides what can be said online.[6] Reduced to its very core, this is the question that this dissertation is about: *who* determines what we can say online, and *how* is that determined? What the DSA does not – and therefore this dissertation does not – address is the more fundamental question of *what* can be said online.

The DSA addresses how internet intermediaries[7] moderate content.[8] Content moderation is a sophisticated process that involves both rule-setting and enforcing those rules. It not only affects whether users can express themselves freely on the internet, but also determines what information is imparted on users.[9] To moderate content, internet intermediaries rely on automated means and human oversight.[10]

Content moderation has implications for the fundamental rights of internet users and for citizens beyond the digital realm.[11] The right most directly affected by content moderation is the right to freedom of expression and to information (in short: the freedom of expression): the freedom of users to express opinions and ideas they think are important, and their freedom to receive information without undue

[5] Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services […], *OJ L 277, 27.10.2022*, p. 1-102

[6] E.g. Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011.

[7] 'Internet intermediaries' is chosen as a neutral term that does not map onto legal definitions such as 'information society service provider', or 'online platform' to avoid excluding a (legal) subcategory of internet service. Although content moderation as a practice involves actors beyond social media platforms (see Chapter II), most societal issues arise from the use of social media platforms. These are however not specified as an individual legal subcategory in the Digital Services Act.

[8] Regulation 2022/2065, art. 3(t): "content moderation means activities, whether automated or not, undertaken by providers of intermediary services that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service, including measures taken that affect the availability, visibility, and accessibility of that illegal content or that information, such as demotion, demonetisation, disabling of access to, or removal thereof, or that affect the ability of the recipients of the service to provide that information, such as the termination or suspension of a recipient's account."

[9] Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 Social Media + Society 205630512211175; Rachel Griffin, 'The Law and Political Economy of Online Visibility: Market Justice in the Digital Services Act' (2023) 2023 Technology and Regulation 69; Natali Helberger, 'On the Democratic Role of News Recommenders' (2019) 7 Digital Journalism 993; Taina Bucher, 'Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook' (2012) 14 New Media & Society 1164.

[10] Ivar A Hartmann, 'A New Framework For Online Content Moderation' (2020) 36 Computer Law & Security Review 1; Robyn Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute 2018); Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 205395172094323.

[11] E.g. Beatriz Botero Arcila and Rachel Griffin, 'Social Media Platforms and Challenges for Democracy, Rule of Law and Fundamental Rights' (European Parliament 2023) PE 743.400.

limitation.[12] Freedom of expression is a core right in democracies, and is generally protected by constitutions and international fundamental rights documents. In the European Union, most crucial for its protection are the European Convention on Human Rights (ECHR)[13], the Charter for Fundamental Rights in the European Union (CFR)[14], and national constitutions.[15] These instruments are addressed Section 2.2.a. Freedom of expression in the European Union is not an absolute right: it may be limited under article 10(2) ECHR and article 52(1) CFR under the condition that the limitation is provided for by law, pursues a legitimate aim, and is necessary and proportionate.[16] In practice, proportionality means that a fair balance must be struck between citizens' right to express themselves and conflicting rights and interests such as non-discrimination, security, public health, or intellectual property rights protection.

This balancing act is a delicate and continuous exercise, and forms one of the cornerstones of content moderation.[17] The fact that this balancing act, usually conducted by state entities, in content moderation is conducted by private actors is cause for debate.[18] Private actors take on a near-sovereign role in the digital realm, and have been difficult to grasp in regulatory instruments. This is partly because regulators are, by default, playing catch-up with internet intermediaries with a significant resource and knowledge advantage,[19] but also because the regulation of content moderation by state actors creates friction with the freedom of expression: regulation may not unduly interfere with the freedom of expression of users.[20] Equally at tension with traditional concepts of freedom of expression is that private entities, through content moderation practices, determine what can be said online although they typically have no direct responsibility for the protection of the

[12] Janneke Gerards, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023) 105.

[13] European Convention on Human Rights, 213 UNTS 222, 4 November 1950.

[14] Charter of Fundamental Rights of the European Union *OJ C 326*, 26 October 2012, p. 391-407.

[15] Generally: Giovanni Pitruzzella and Oreste Pollicino, 'The Constitutional Perspective on Freedom of Expression in the Internet Era' in Andrew Spannaus (tr), *Disinformation and hate speech: a European constitutional perspective* (First edition, Bocconi University Press 2020); Lorna Woods, 'Digital Freedom of Expression in the EU' in Sionaidh Douglas-Scott and Nicholas Hatzis, *Research Handbook on EU Law and Human Rights* (Edward Elgar Publishing 2017); Wolfgang Benedek and Matthias C Kettemann, 'Freedom of Expression and the Internet' (Council of Europe 2013).

[16] E.g.: European Court of Human Rights, 26 April 1979, *The Sunday Times v the United Kingdom*, 6538/74, para 49; European Court of Human Rights, 2 April 2013, *Animal Defenders International v the United Kingdom*, 48876/08, para 100.

[17] Kyle Langvardt, 'Regulating Online Content Moderation' (2017) 106 The Georgetown Law Journal 1353, 1363.

[18] E.g. Tarleton Gillespie, *Custodians of the Internet* (1st edn, Yale University Press 2018); Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598.

[19] Robert Gorwa, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024) 66–69; Gillespie, *Custodians of the Internet* (n 18).

[20] E.g. Balkin (n6), 2017.

freedom of expression of their internet users.[21] Their commercial interests may put freedom of expression under pressure, for example when unpopular expressions are removed to benefit advertising revenue,[22] or when politically sensitive speech is removed as part of keeping the platform accessible to everyone.[23]

The DSA also addresses tensions with fundamental rights in the provision of intermediary services and content moderation.[24] The DSA is part of a wide range of regulations that attempts to limit power of internet intermediaries. It provides a horizontal, procedural framework that aims to instil responsibility and accountability with providers of intermediary services. Amongst different regulatory interventions, it formalises the roles and interactions of a range of stakeholders in the content moderation process that have influence on the scope of freedom of expression on the internet:[25] users, NGOs, trusted flaggers, out-of-court dispute settlement bodies, and national regulators are all attributed a role in the governance of internet intermediaries.[26]

The DSA's formalisation of a range of stakeholders in the content moderation process raises questions from a fundamental rights perspective, as content moderation largely determines the scope of the freedom of expression on the internet. The formalisation of each individual stakeholder in content moderation affects the scope of freedom of expression on the internet: regulators may enforce national laws across the EU and display biases in their enforcement of the DSA;[27] platforms may shape how users interact with content moderation through reporting mechanisms;[28] and civil society organisations may represent single, specific interests

[21] Balkin (n 6); Klonick, 'The New Governors' (n 18). See infra Section II.ii.c.

[22] Rachel Griffin, 'From Brand Safety to Suitability: Advertisers in Platform Governance' (2023) 12 Internet Policy Review.

[23] E.g. 'Fury over Facebook "Napalm Girl" Censorship' *BBC News* (9 September 2016) <https://www.bbc.com/news/technology-37318031> accessed 27 August 2025.

[24] Broadly: Martin Husovec and Irene Roche Laguna, 'Digital Services Act: A Short Primer' [2022] SSRN Electronic Journal <https://www.ssrn.com/abstract=4153796> accessed 7 September 2022; Folkert Wilman, 'The Digital Services Act (DSA): An Overview'.

[25] E.g. Rachel Griffin, 'The Politics of Risk in the Digital Services Act: A Stakeholder Mapping and Research Agenda' (2025) 5 Weizenbaum Journal of the Digital Society; Niklas Eder, 'Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation' [2024] German Law Journal 1; Jörg Wimmers, 'The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act' (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 381.

[26] See for an overview Andrej Savin, 'The EU Digital Services Act: Towards a More Responsible Internet' (2021) 24 Journal of Internet Law 15.

[27] E.g. Jennifer Orlando-Salling, 'The Digital Services Act in the European Periphery: Critical Perspectives on EU Digital Regulation' [2025] European Law Open 1.

[28] Marie-Therese Sekwenz, Ben Wagner and Simon Parkin, '"It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!" – Evaluating Reporting Mechanisms under the Digital Services Act', *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2025).

that may not take the freedom of expression into account in their interactions with content moderation.[29]

These tensions warrant a further exploration of the DSA's formalisation of these actors in the content moderation process in light of their impact on freedom of expression. This is especially pertinent since content moderation and the power structures surrounding it have been critiqued from all corners of an increasingly polarised society. Clarification of such issues may inform the debate and foster consensus on how to approach internet regulation. Central to this dissertation is the argument that the formal involvement of a variety of actors in the content moderation process changes the balancing of the right to freedom of expression against other interests. This intervention is not exhaustively resolved by the Digital Services Act. In six articles, this dissertation addresses the formal role in content moderation created for several actors in the DSA, and explores how that intervention affects balancing the freedom of expression against other interests. Where appropriate, suggestions are made for safeguards to ensure that unjustified or unnecessary risks for the freedom of expression of internet users are minimised, or conversely, that the freedom of expression needs to be more carefully balanced against other interests. In doing so, this dissertation explores the following overarching research question:

How do the Digital Services Act and the formal roles therein for internet intermediaries, users, civil society, and state actors affect the balancing of the right to freedom of expression of internet users and other interests in content moderation?

The research question is divided into sub-questions concerning three main categories of actors, namely internet intermediaries, civil society and users, and state actors. This dissertation does not address all actors that fall under the DSA. In case of internet intermediaries, it focuses predominantly on hosting providers in the sense of article 3(g)(iii) DSA. Social media platforms are the most societally impactful subset thereof. However, findings in this dissertation indicate that content moderation by other hosting providers is equally relevant and must also be considered. For civil society, the dissertation looks primarily at trusted flaggers. For state actors, both state and supra-state actors are considered, with a specific focus on digital services coordinators.

[29] Brenda Dvoskin, 'Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance' (2022) 67 Villanova Law Review 447.

This introductory chapter proceeds by providing the necessary background to answer these research questions. It sets out the evolution of EU regulation of internet intermediaries into the DSA. It also discusses the role of freedom of expression within the DSA and content moderation in general, and it explains why the DSA can be seen as a formalisation of networked platform governance. Networked platform governance is a theory developed by Robyn Caplan and describes multistakeholderism in platform governance.[30] The theory can be used to describe the interactions the DSA facilitates between stakeholders that affect platform governance. To understand networked platform governance the introductory chapter describes what the general theory of multistakeholderism encompasses, and how networked platform governance applies multistakeholderism to platform governance. Following this background sketch, it introduces the outline of the dissertation and the methodology by which the research questions are addressed.

2. Background

2.1. From e-Commerce Directive to Digital Services Act[31]

The regulation of information society services has been a puzzle that regulators have sought to solve since the 1990s. The challenges that information society services pose are not always sufficiently met by existing laws for the 'offline' world, and may require different regulatory approaches. The latest piece in the puzzle on internet regulation is the European Union's Digital Services Act. The DSA is a novel regulatory instrument that introduces a wide range of transparency and due diligence obligations that internet intermediaries[32] must respect on top of the existing liability exemptions that are based on the earlier e-Commerce Directive (Directive 2000/31, 'ECD').[33] It also formalises the role of a number of actors in content moderation: users, platforms, civil society, and state actors. The DSA, arguably more than preceding regulations, emphasises the protection of fundamental rights in the digital realm. However, some of its novel regulatory approaches can create friction with those rights; the DSA depends to a considerable degree on self-and co-

[30] Robyn Caplan, 'Networked Platform Governance: The Construction of the Democratic Platform' (2023) 17 International Journal of Communication 3451.

[31] Some of the analyses and backgrounds in this subsection are also presented in Jacob van de Kerkhof and Catalina Goanta, 'Van Code Is Wet Naar Wet Is Code: Geautomatiseerde Transparantie in de Digitaledienstenverordening' [2025] Tijdschrift voor Internetrecht 19.

[32] The Regulation applies to 'intermediary services', and often refers to 'information society services' as defined in Directive 2015/1535, art. 1(1)(b). See *supra* footnote 7.

[33] 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 July 2000, p 1-16; see *infra* Chapter III.

regulation, meaning that safeguarding and balancing fundamental rights will be done partially by private actors.[34] In addition to that, it uses relatively open-ended definitions in its systemic risk framework, which lack clear empirical indicators in their assessment.[35] While these approaches increase the future-proofness of the regulation by covering potential risks stemming from currently undeveloped services to be included, they create a degree of tension with the principles of rule of law and legal certainty.

To illustrate how the evolution of internet regulation has resulted in the DSA, and how this particular piece of internet regulation interacts with fundamental rights, this subsection consists of two parts. The first part offers an introduction to the ECD and its discontents, with a focus on the issues that led to adopting the Digital Services Act. This part highlights the conundrum that internet regulation creates from a fundamental rights perspective, showing that regulating the internet is a delicate balancing act between over- and under-regulation. The second part describes the Digital Services Act, and highlights some of the provisions that will be the focal point of the following chapters.

2.1.a. An introduction to the e-Commerce Directive and its discontents

At the advent of internet regulation, hosting content on the internet posed primarily liability questions, and not necessarily questions about large scale societal harms caused by systemic impacts of online services. The ECD is the EU's first legislation addressing this liability question for information society service providers. 'Information society services', defined in Directive 98/48, are defined as 'any service that is normally provided for by remuneration at a distance, by electronic means, and at the individual request of a recipient of services.'[36] The ECD makes a distinction between mere conduit, caching, and hosting services.[37]

Content online may infringe on intellectual property, contain abusive material, or be otherwise illegal. This raises the question of who can be held liable for that illegality: the person posting content, who may be an anonymous internet user, the person

[34] Giovanni De Gregorio and Pietro Dunn, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473.

[35] Kester Siegrist, *Systemische Risiken im Digital Services Act* (1. Auflage, Nomos 2025) 128–129.

[36] Directive 1998/48 of the European parliament and of the Council of 20 July 1998 amending Directive 98/34/EC laying down a procedure for the provision of information in the field of technical standards and regulation, *OJ L 217*, 5 August 1998, p. 18-26. art. 1(2); later codified in Directive 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations of rules on Information Society Services, *OJ L 241*, 17 September 2015, art. 1(1)(b).

[37] Directive 2000/31, art. 12-14.

whose page that content is posted on,[38] the online platform operator that hosts the content,[39] the domain-name host facilitating the platform's presence on the web,[40] or even the party facilitating internet access to that illegal content.[41] Each of these actors has a reason for desiring to be exempted from liability for their role in storing illegal content. Edwards describes how their plea for immunity generally rests on three arguments: "Lack of effective legal or actual control [over the illegal content]; the inequity of imposing liability upon a mere intermediary ('shooting the messenger'); and in Europe especially, consequences for the public interest if the unlimited liability was, nonetheless, imposed."[42] With the last argument Edwards refers to potential harms to freedom of expression if liability leads to over-removal of content.

Naturally, arguments for complete exemption from liability need to be weighed against the interests of parties aggrieved by illegal content. The ECD weighs these interests by making liability exemption conditional on the internet intermediary having knowledge of the illegal content and failing to remove it. This symbolises a trade-off that ensures that illegal content can be removed when a hosting provider is requested to do so by aggrieved parties, while maintaining the expected social gains such as freedom of expression and economic benefits associated with limited liability. Part of this trade-off is the notice-and-action mechanism.[43] When users submit a notice signalling to the internet intermediary that content is illegal, the internet intermediary is expected to remove that content expeditiously after confirming that the content is indeed illegal and removing is it compatible with the freedom of expression.[44] However, aside from this core obligation, the ECD did little to incentivise responsibility and accountability from internet intermediaries for the effects of their services.

The ECD predates developments in the online landscape such as social media platforms, automation, and commercialisation of the online sphere. The ECD was adopted in 2000, at the beginning of a development dubbed 'Web 2.0'. Web 2.0

[38] See e.g. European Court of Human Rights, 15 May 2023, *Sanchez v France*, 45581/15.

[39] See e.g. Court of Justice of the European Union, 22 June 2021, *Peterson v Google/Elsevier v Cyando*, C-682/18 & C-683/18.

[40] See e.g. Swedish District Court, 19 May 2015, *.SE*, see more extensively M Truyens and P Van Eecke, 'Liability of Domain Name Registries: Don't Shoot the Messenger' (2016) 32 Computer Law & Security Review 327, specifically fn 33 case law cited therein.

[41] Court of Justice of the European Union, 15 September 2016, *Tobias McFadden v Sony Music Entertainment*, C-484/14.

[42] Lilian Edwards, *Law, Policy, and the Internet* (Hart publishing 2019) 257.

[43] Generally: Aleksandra Kuczerawy, 'From "Notice and Takedown" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020); Christina Angelopoulos and Stijn and Smet, 'Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability' (2016) 8 Journal of Media Law 266.

[44] Court of Justice of the European Union, 22 June 2021, *Peterson v Google/Elsevier v Cyando*, C-682/18 & C-683/18, para 116.

signifies a change in web usage, "whereby content and applications are no longer created and published by individuals, but instead continuously modified by all users in a participatory and collaborative fashion."[45] Collaborative platforms and participatory culture increased the opportunities for the public on the web exponentially.[46] This is also visible in terminology used to describe such developments online. For example, early literature speaks of social networking sites, which emphasises a networking aspect. Such networking sites allow individuals to "construct a public or semi-public profile within a bounded system, articulate a list of users with whom they share a connection, and view and transverse their list of connections and those made by others within the system."[47] By contrast, later literature often uses the term 'social media platforms' which better reflects the role of the platform as an aggregator of content with a wide range of mechanisms and functions.[48]

It is beyond the scope of this dissertation to provide an exhaustive history of online platforms, but it is clear that these now form a ubiquitous part of the life of billions of people. Especially social media platforms have become complex software infrastructures that facilitate interaction through predominantly picture and film-focused communications.[49] Online platforms have become networking fora, entertainment systems, news outlets, and marketplaces, all at the same time. A fitting analogy is J.K. Rowling's Mirror of Erised: a magical mirror that shows the deepest, most desperate desire of our hearts.[50] The content that is presented by online platforms seamlessly aligns with our interests: a powerful attribute that risks causing significant harms. Harms from online platforms can manifest on an individual level, for example through addictive behaviour[51] or mental health harms,[52] but online platforms can also have deep impact on society, such as by

[45] Andreas M Kaplan and Michael Haenlein, 'Users of the World, Unite! The Challenges and Opportunities of Social Media' (2010) 53 Business Horizons 59, 61.

[46] José van Dijck, Thomas Poell and Martijn de Waal, *The Platform Society: Public Values in a Connective World* (Oxford University Press 2018) 2.

[47] Danah M Boyd and Nicole B Ellison, 'Social Network Sites: Definition, History, and Scholarship' (2007) 13 Journal of Computer-Mediated Communication 210, 211.

[48] Dijck, Poell and Waal (n 46) 31–48.

[49] Vibhor Agarwal and Nishanth Sastry, '"Way Back Then": A Data-Driven View of 25+ Years of Web Evolution', *Proceedings of the ACM Web Conference 2022* (ACM 2022) <https://dl.acm.org/doi/10.1145/3485447.3512283> accessed 5 September 2022.

[50] J.K. Rowling, 'Harry Potter and the Philosopher's Stone', Chapter 12: The Mirror of Erised; the mirror "shows the most desperate desire of a person's heart, a vision that has been known to drive men mad."

[51] E.g. Yubo Hou and others, 'Social Media Addiction: Its Impact, Mediation, and Intervention' (2019) 13 Cyberpsychology: Journal of Psychosocial Research on Cyberspace <https://cyberpsychology.eu/article/view/11562> accessed 25 July 2025.

[52] E.g. John A Naslund and others, 'Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice' (2020) 5 Journal of Technology in Behavioral Science 245.

affecting election outcomes,[53] they can cause societal shifts,[54] and they have even been associated with sparking genocide.[55]

The development of online platforms and its associated harms were scarcely covered by the ECD, which predominantly provides a framework that addresses individual pieces of illegal content, but not the profound systemic impact and adverse effects to fundamental rights that platform practices can have. The ECD was ill-adjusted to the new reality of large online platforms in at least three – interrelated – respects: there was no harmonised procedural framework around notice-and-action mechanisms, which form the core of the liability exemption in the ECD;[56] it provided few safeguards to fundamental rights in the digital realm, including the freedom of expression;[57] and it did not clarify what responsibilities internet intermediaries have at a systemic level,[58] beyond generally being a 'diligent economic operator'.[59] These shortcomings led to national initiatives such as the German *Netzwerkdurchsetzungsgesetz* (*NetzDG*)[60], the French *Loi Avia*[61], and the Austrian *Hass-im-Netz-Bekämpfungsgesetz*[62], which fragmented the regulatory landscape for internet intermediaries. These national laws were not uncontroversial; *NetzDG* attracted wide attention for potentially 'chilling' the freedom of expression,[63] and *Loi Avia* was even

[53] E.g. in the USA: Sam Meredith, 'Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal' (*CNBC*, 10 April 2018) <https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html> accessed 18 July 2025; the United Kingdom: Max Hänska and Stefan Bauchowitz, 'Tweeting for Brexit: How Social Media Influenced the Referendum' in John Mair and others (eds) (Abramis Academic Publishing 2017) <http://www.abramis.co.uk/publish/home.php> accessed 25 July 2025; or Romania: 'Romanian Authorities Ask EU to Investigate TikTok Role in Election' (*euronews*, 13:47:26 +01:00) <https://www.euronews.com/next/2024/11/27/romanian-authorities-ask-eu-commission-to-investigate-tiktok-role-in-election> accessed 18 December 2024;

[54] For example, the Arab Spring: Gadi Wolfsfeld, Elad Segev and Tamir Sheafer, 'Social Media and the Arab Spring: Politics Comes First' (2013) 18 The International Journal of Press/Politics 115. And the Black Lives Matter movement: Ho-Chun Herbert Chang, Allissa Richardson and Emilio Ferrara, '#JusticeforGeorgeFloyd: How Instagram Facilitated the 2020 Black Lives Matter Protests' (2022) 17 PLOS ONE e0277864.

[55] Barbara Ortutay and Associated Press, 'Amnesty Report Finds Facebook Amplified Hate Ahead of Rohingya Massacre in Myanmar' (*PBS News*, 29 September 2022) <https://www.pbs.org/newshour/world/amnesty-report-finds-facebook-amplified-hate-ahead-of-rohingya-massacre-in-myanmar> accessed 18 July 2025.

[56] Edwards (n 42) 270–277.

[57] E.g. European Parliament resolution of 20 October 2020 on the Digital Services Act and fundamental rights issues posed (2020/2022(INI)), accessible via: http://europarl.europa.eu/doceo/document/TA-9-2020-0274_EN.html

[58] Martin Husovec, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation Of Global Internet Rules' (2023) 38 Berkeley Technology Law Journal 883, 897–898. *Cf.* Directive 2000/31, rec. 48.

[59] Court of Justice of the European Union, 22 June 2021, *Youtube/Cyando*, C-682/18 and C-683/18, para 116; More in-depth see *infra* Chapter III.

[60] *Netzwerkdurchsetzungsgezetz*, 1 September 2017, BGBl I 2959.

[61] *Loi du 24 Juin 2020 visant à lutter contre les contenus haineux sur internet*, 24 June 2020, 2020-766.

[62] *Hass-im-Netz-Bekämpfungs-Gesetz*, 18 November 2020, Bundesgesetzblatt I Nr. 148/2020.

[63] Robert Gorwa, 'Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG' (2021) 45 Telecommunications Policy 102145, 7.

deemed unconstitutional by the French Constitutional Council.[64] Moreover, the fragmented landscape created uncertainty for internet intermediaries in the European Union. The DSA seeks to resolve this fragmentation by harmonising requirements for notice-and-action and protection of fundamental rights and creating clearer expectations for internet intermediaries.

2.1.b. The Digital Services Act

The DSA is part of the EU's 2020 Digital Strategy that seeks to adopt new rules "to deepen the Internal Market for Digital Services by increasing and harmonising the responsibilities of online platforms and information service providers."[65] At its core it is a market harmonisation instrument based on article 114 TFEU.[66] It is a full harmonisation instrument that precludes national regulation going beyond the limits of the DSA.

The DSA addresses the regulatory vacuum left by the ECD by clarifying responsibilities for internet intermediaries and addressing fundamental rights concerns that arose from the centralisation of content moderation power. It does so by providing a procedural framework for internet intermediaries to follow in providing their services that seeks to minimise the risk of harms emanating from these services. It is a consolidation of various policy options discussed in the European Commission's Impact Assessment, which aimed to explore the changed nature of digital services.[67] The DSA includes *inter alia* the codification of the 2018 Recommendation on Illegal Content,[68] which created a stronger framework for intermediary liability exemption and transparency measures.

Broadly speaking the DSA has four functions: (i) it strengthens the previous notice-and-action regime for platform liability with procedural safeguards, including for oversight authorities; (ii) it creates a range of due diligence obligations and transparency measures for internet intermediaries that increase as the scope of their service increases; (iii) it creates an enforcement network consisting of national authorities and the European Commission; (iv) it formalises a range of actors in the

[64] *Conseil Constitutionnel*, 18 June 2020, *Loi visant à lutter contre les contenus haineux sur internet*, 2020-801 DC.

[65] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM(2020)825 final, Explanatory Memorandum.

[66] Treaty on the Functioning of the European Union, *OJ C* 202, 7.6.2016, p. 1-388.

[67] 'Impact Assessment of the Digital Services Act | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act> accessed 18 July 2025. See Savin (n 24) 19.

[68] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online.

content moderation process as part of the due diligence obligations for internet intermediaries. The following paragraphs explain this in more depth.

First, the DSA adopts the ECD's intermediary liability exemption provisions almost *verbatim* in articles 4-6.[69] The DSA also differentiates between mere conduit, caching, and hosting services. The balance struck in the conditional liability exemption between freedom of expression, interests of aggrieved parties, and economic interests of internet intermediaries was deemed 'fit for purpose' in the Commissions preparatory works.[70] There was no support to depart from these core principles, especially in light of sectoral regulation meeting demands for stricter liability rules in specific industries.[71] The framework is strengthened with procedural norms around notice-and-action,[72] codification of case law in the DSA's recitals,[73] and norms for takedown orders by oversight authorities.[74]

Second, a novelty of the DSA lies in provisions that create a layer of accountability for how platform services operate.[75] Accountability is ensured through a range of transparency obligations through which regulators and civil society can hold internet intermediaries accountable for their practices, and due diligence obligations that seek to ensure that platforms critically evaluate their practices and address harmful aspects.[76] The accountability mechanisms are based on a risk-based approach, meaning that most enforcement activities will focus on factors from which most risks arise.[77] The concept of risk in the DSA is defined non-exhaustively along four categories: (i) dissemination of illegal content; (ii) negative effects for the exercise of fundamental rights; (iii) negative effects on civic discourse and electoral processes; (iv) negative effects on physical or mental well-being.[78] The indeterminate nature of this concept has sparked rich debate, which is left relatively unaddressed

[69] See more on this *infra* Chapter III.

[70] Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) 94.

[71] E.g. João Pedro Quintais and Sebastian Felix Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?' (2022) 13 European Journal of Risk Regulation 191.

[72] Regulation 2022/2065, art. 16; see *infra* Chapter III and IV.

[73] See on this: Ronan Riordan, 'A Case Study of Judicial-Legislative Interactions via the Lens of the DSA's Host Liability Rules' (2025) 2025 10 European Papers - A Journal on Law and Integration 259.

[74] Regulation 2022/2065, art. 9.

[75] Generally Husovec (n 58).

[76] Husovec (n 70) 182–184.

[77] Claudia Quelle, 'Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach' (2018) 9 European Journal of Risk Regulation 502; De Gregorio and Dunn (n 34).

[78] Regulation 2022/2065, art. 34(1), see also Ruth Janal, 'Haftung Und Verantwortung Im Entwurf Des Digital Services Acts' [2021] Zeitschrift für Europäisches Privatrecht 227, 266.

in this dissertation.[79] Part of the accountability framework are extensive transparency requirements.[80] Internet intermediaries are required to be transparent over terms and conditions,[81] content moderation decisions towards users and other stakeholders,[82] and are required to fulfil a range of reporting obligations,[83] partake in audits,[84] and facilitate data access for researchers and oversight authorities.[85] Transparency forms a cornerstone in creating accountability, but also in facilitating enforcement.[86]

Due diligence obligations and transparency measures depend on the size and type of the intermediary service. The DSA as a whole applies to 'information society services', a definition borrowed from article 1(1)(b) of Directive 2015/1535, that refers to any service normally provided for remuneration, at a distance, by electronic means, and at the individual request of a recipient of services. 'Hosting services'[87] are a subset of that definition, that, among other tasks, have to facilitate notice-and-action mechanisms and statements of reasons. 'Online platforms' are in turn a subset of hosting providers.[88] These service providers are subject to obligations on transparency reporting, complaint handling, out-of-court dispute settlement, trusted flaggers, and product design requirements on recommender systems and dark

[79] It is unclear at what point a risk is significant enough to be 'systemic', what aspects of platform services are covered, and how the obligation to mitigate systemic risks arising from legal content ('awful but lawful') relates to freedom of expression and the rule of law. See e.g. Rachel Griffin, 'Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act' [2025] European Law Open 1; Mateus Correia de Carvalho, 'It Will Be What We Want It to Be: Sociotechnical and Contested Systemic Risk at the Core of the EU's Regulation of Platforms' AI Systems' [2025] Journal of Intellectual Property, Information Technology and Electronic Commerce Law 35; Katharina Kaesling and Annegret Wolf, 'Sustainability and Risk Management under the Digital Services Act: A Touchstone for the Interpretation of "Systemic Risks"' (2025) 74 GRUR International 119; Sally Broughton Micova and Andrea Calef, 'Elements for Effective Systemic Risk Assessment Under the DSA' (Centre on Regulation in Europe 2023); Agustina Del Campo, Nicolas Zara and Ramiro Álvarez-Ugarte, 'Are Risks the New Rights? The Perils of Risk-Based Approaches to Speech Regulation' (2025) 16 Journal of Intellectual Property, Information Technology and Electronic Commerce.

[80] See generally van de Kerkhof and Goanta (n 31).

[81] Regulation 2022/2065, art. 14(1).

[82] Regulation 2022/2065, art. 17(1); art. 24(5).

[83] Regulation 2022/2065, art. 15; 24; 39; 42.

[84] Regulation 2022/2065, art. 37(1)

[85] Regulation 2022/2065, art. 40(1) & (4).

[86] Commission Delegated Regulation (EU) of 1 July 2025 supplementing Regulation 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data with vetted researchers, rec. 2; see also https://digital-strategy.ec.europa.eu/en/faqs/dsa-transparency-database-questions-and-answers.

[87] Regulation 2022/2065, art. 3(g)(iii): "a 'hosting' service, consisting of the storage of information provided by, and at the request of, a recipient of the service".

[88] Regulation 2022/2065, art. 3(i); "'online platform' means a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the applicability of this Regulation."

patterns. The most stringent regime in the DSA applies to Very Large Online Platforms and Very Large Online Search Engines ('VLOPSEs).[89] These service providers are subject to a range of transparency, risk assessment, mitigation, and auditing obligations, as they are perceived to bring the most risk.[90] Figure 1 outlines the layered approach.[91]

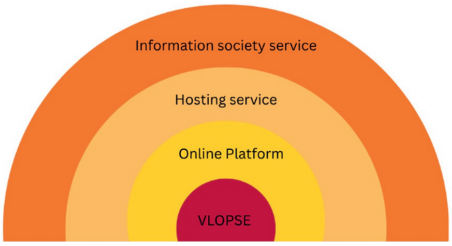


Figure 1: the DSA's layered approach

Third, the DSA introduces a strengthened oversight framework. This framework consists of a combination of national authorities, the European Commission, and input from independent auditors and researchers.[92] The DSA introduces a harmonised procedure for competent national supervisory authorities to issue takedown orders.[93] Supervisory authorities also have a role in accrediting stakeholders in content moderation, such as out-of-court dispute settlement bodies and trusted flaggers.[94] This network allows oversight authorities to sanction non-compliance from internet intermediaries by fines, binding agreements on changing aspects of the intermediary service, or even blocking certain aspects of the hosting provider's service. The oversight network is a hybrid between a new national authority, the Digital Services Coordinator[95] and the European Commission.[96] In

[89] Regulation 2022/2065, art. 33(1): "online platforms and online search engines which have a number of average monthly active recipients of the service in the Union equal to or higher than 45 million, and which are designated as very large online platforms or very large online search engines pursuant to paragraph 4."
[90] Regulation 2022/2065, rec. 48.
[91] Inspired by Quintais and Schwemer (n 71) 18 figure 4.
[92] Eder (n 25); Anna Liesenfeld, 'The Legal Significance of Independent Research Based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act' [2024] European Journal of Risk Regulation 1.
[93] Regulation 2022/2065, art. 9.
[94] Regulation 2022/2065, arts. 21 & 22.
[95] Regulation 2022/2065, art. 49(1). See specifically *infra* Chapter VI.
[96] Regulation 2022/2065, art. 56(3).

addition to public authorities sanctioning platforms for non-compliance, the DSA also facilitates possibilities for private litigation for non-compliance.[97]

Fourth, and relating to the previous point; the DSA formalises the roles of several actors in platform governance and facilitates interactions between them. At the very core of this framework remain the internet intermediaries. The DSA limits their power in content moderation to some extent. For example, article 16 sets out a clear procedure for hosting providers to follow when a user submits a notice on illegal content, which, if followed and if the content is deemed illegal by the hosting provider, requires the hosting provider to remove that content or risk liability.[98] To some extent this still creates a dependency relationship between user and hosting provider. This dissertation explores in Chapter IV how hosting providers facilitate the submission of notices determines the effectiveness of the user's role in content moderation. In spite of the proceduralisation, platforms have a relatively wide discretion. The process of treating notices is largely an obligation of effort, requiring solely that notices are treated in a timely and impartial manner.[99] When the procedure is followed, platforms must provide users with a statement of reasons.[100] Such statements of reasons can lead to an appeal: users can complain with the platform, and the online platform must hear their complaints and respond without undue delay.[101]

When a user chooses not to lodge a complaint on the content moderation decision with the platform, they may still appeal the content moderation decision at an out-of-court dispute settlement body. These institutions mimic earlier efforts of 'judicialising' content moderation such as the Meta Oversight Board.[102] Such bodies are intended to provide fair, swift, and cost-effective dispute resolution for recipients of platform services.[103] They simultaneously form an avenue for an independent interpretation of platform community guidelines.[104] The DSA also provides a formal

[97] Paddy Leerssen and others, 'Pathways to Private Enforcement of the Digital Services Act (DSA)' (Institute for Information Law 2025). See also Jacob van de Kerkhof and Catalina Goanta, 'Shadowbanned on X: The DSA in Action' [2024] European Journal of Risk Regulation 1.

[98] Regulation 2022/2065, art. 16(3).

[99] Daniel Holznagel, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (2024) 25 Computer Law Review International 172.

[100] Regulation 2022/2065, art. 17(1).

[101] Regulation 2022/2065, art. 20(4).

[102] Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 Yale Law Journal 2418.

[103] Regulation 2022/2065, art. 21(1).

[104] Pietro Ortolani, 'If You Build It, They Will Come' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-build-it/> accessed 30 September 2024.

procedure for trusted flaggers in content moderation, who can submit notices that are privileged over 'ordinary' notices.[105] The most notable shift here is that platforms may no longer choose not to interact with these stakeholders; engaging with various stakeholders is legally required by the DSA. A similar anchoring is visible in the establishment of codes of conduct.[106] Even though these were previously voluntary collaborations between various stakeholders, their regulatory inclusion creates the expectation of participation by internet intermediaries, especially since abstention from these codes of conduct is observed as non-compliance.[107]

The DSA's proceduralisation of content moderation and the roles of various actors therein raises questions about the freedom of expression of internet users. As established earlier, content moderation is inseparably linked to freedom of expression.[108] The next subsection explains the tensions content moderation creates with the freedom of expression.

2.2. Freedom of Expression in the Digital Realm

The freedom of expression is an important concept in this dissertation. It serves as a point of reflection that is used to evaluate the impact of the DSA's involving different actors in content moderation. This subsection explains the framework for protection of freedom of expression in the European Union, discusses how this framework applies to internet intermediaries, and how it relates to content moderation.

2.2.a. The European Convention on Human Rights and the Charter for Fundamental Rights of the European Union

As mentioned in the introductory section, freedom of expression in the EU is protected by a range of instruments: national constitutions, the ECHR, and the CFR. It is beyond the scope of this introduction to address how each of the 27 national constitutions protect the right to freedom of expression. Instead, the dissertation focuses on the protection of freedom of expression in the ECHR and CFR as the leading instruments regarding fundamental rights in the EU. The ECHR is a document pertaining to the legal order of the Council of Europe. As such, it has more signatories than just the EU member states – however, all EU member states

[105] Regulation 2022/2065, art. 22, see Chapter V

[106] Regulation 2022/2065, art. 45-47. See specifically *infa* Chapter VII.

[107] Regulation 2022/2065, rec. 104; see also Rachel Griffin, 'Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns' (2024) 2024 Technology and Regulation 167.

[108] E.g. Joan Barata, 'The Future of Free Speech: Old Threats and New Challenges' in Eric Heinze and others (eds), *Oxford Handbook of Hate Speech* (Oxford University Press forthcoming) 3.

are party to the ECHR, and the Convention and the Charter overlap thus *ratione personae*. The EU as such has the intention to accede the Convention, but has not done so yet.[109] The CFR is applicable to the EU as a whole,[110] and to each of its member states individually when they implement EU law. The CFR forms the guiding document pertaining to fundamental rights with regard to European regulation, and is referenced throughout the DSA.

Both documents build on a similar core of fundamental rights protection in Europe, and as such also overlap *ratione materiae.* The CFR contains 'more' fundamental rights than the ECHR; since the ECHR predates the CFR by some 50 years, the CFR could include specific provisions reflecting the present-day understanding of fundamental rights. For example: data protection under the ECHR is protected under the right to private life in article 8,[111] whereas in the CFR it is a self-standing right under article 8.[112] However, many provisions correspond to a significant extent. The substantive correspondence between the two documents could potentially lead to diverging interpretations, since the documents are interpreted by different courts: the ECHR is interpreted by the European Court of Human Rights (ECtHR), whereas the Charter is interpreted by the Court of Justice of the European Union (CJEU).[113]

The risk of divergence in the CJEU interpretation of the CFR in relation to the ECHR is partially resolved through article 52(3) CFR, which stipulates that in the case of corresponding rights, the rights of the Charter must be interpreted in light of the European Convention, and therefore the case law of the ECtHR.[114] A similar solution has been found for the risk that EU member states could experience a conflict between their obligations as EU member state and those as Convention member state.[115] In cases where the obligations following from EU law[116] or procedure[117] are

[109] Treaty on European Union, art. 6(2); see also Court of Justice of the European Union, 18 December 2014, Opinion 2/13.

[110] Treaty on European Union, art. 6(1).

[111] E.g. European Court of Human Rights, 4 December 2008, *S. and Marper v the United Kingdom,* 30562/04 & 30566/04, para 67.

[112] Charter of Fundamental Rights of the European Union, art. 8. The right to private life is codified in article 7.

[113] Maribel González Pascual, 'A European Standard of Human Rights Protection?' in Sara Iglesias Sánchez and Maribel González Pascual (eds), *Fundamental Rights in the EU Area of Freedom, Security and Justice* (1st edn, Cambridge University Press 2021) 39-41.

[114] Charter of Fundamental Rights of the European Union, art. 52(3). It is important to note that this interpretation does not prevent the Charter providing a more extensive protection than the ECHR.

[115] See generally Šejla Imamović, 'A European Web of Human Rights Regimes: The EU, the ECHR and the National Legal Systems', *The Architecture of Fundamental Rights in the European Union* (Hart Publishing 2022).

[116] European Court of Human Rights, 18 February 1999, *Matthews v United Kingdom,* 24833/94

[117] European Court of Human Rights, 20 January 2009, *PO Kokkelvisserij v the Netherlands,* 13645/05.

non-discretionary,[118] member states are deemed to have acted in compliance with the Convention when they acted through an international organisation that is considered to protect fundamental rights in a manner that is equivalent to that for which the Convention provides, unless that protection is manifestly deficient.[119] In any instance where member states do have discretion, Convention states must respect the ECHR.[120] To reconcile EU law with Convention law, the ECtHR may align its interpretations of the Convention with EU law, for example when an EU regulation gives expression to rights and freedoms in the Convention;[121] however, it does not have to.[122] In practice, the Courts refer to each other's case law.

In sum, the case law of the ECtHR as it has developed over the past decades is guiding in the interpretation of corresponding provisions contained in instruments. The next subsections will therefore predominantly address ECtHR case law about the freedom of expression. Subsection 2.2.b explains how the freedom of expression is protected in the Convention and the Charter, subsection 2.2.c explains the obligations for state actors for the protection of freedom of expression, and subsection 2.2.d discusses how it affects private conduct by internet intermediaries.

2.2.b. The Protection of Freedom of Expression

Both the Convention and the Charter protect the right to freedom of expression, in articles 10 and 11 respectively. The freedom of expression of article 11 CFR must be interpreted similarly to article 10 ECHR.[123] The right to freedom of expression includes the freedom to hold opinions, and to receive and impart information, without interference by a public authority.[124] It is important to note that this is a two-sided right: it encompasses the right to express, but also the right to receive information.[125] This freedom rests on two bases: expressing what one feels is

[118] Although the DSA is a full harmonisation instrument that strives towards little divergence between member states (see Regulation 2022/2065, rec. 9), member states have discretion over their enforcement actions.

[119] European Court of Human Rights, 30 June 2005, *Bosphorus v Ireland,* 45036/98, paras 155-156.

[120] E.g. European Court of Human Rights, 21 November 2019, *Ilias and Ahmed v Hungary,* 47287/15, para 97.

[121] For example, in European Court of Human Rights, 8 July 2025, *Google v Russia,* 37027/22, para 50, the ECtHR reflects on the DSA as an instrument that balances interests when governments order the removal of content, see more specifically Judge Pavli's concurring opinion.

[122] e.g. Oreste Pollicino, 'European Judicial Dialogue and the Protection of Fundamental Rights in the New Digital Environment: An Attempt at Emancipation and Reconciliation' in Sonia Morano-Foadi and Lucy Vickers (eds), *Fundamental Rights in the EU: A Matter for Two Courts* (Bloomsbury Publishing 2015), 107-108.

[123] Court of Justice of the European Union, 4 May 2016, *Philip Morris,* C-547/14, para 147.

[124] European Convention on Human Rights, art. 10(1).

[125] European Court of Human Rights, 13 February 2003, *Cetin et al v Turkey,* 40153/98 and 40160/98, para 64.

important is an essential aspect of personal autonomy and human dignity;[126] moreover, freedom of expression is necessary to effectively participate in public debates and therefore for democracy as a whole.[127] The freedom of expression in the EU is not absolute, but it should in principle have a wide scope. Expressions may "offend, shock, or disturb the state or any sector of the population.[128] This freedom applies equally offline as it does online.[129] Some types of speech may be excluded from protection. For example, speech that does not express an opinion or impart information but has the sole intention to insult,[130] speech that stirs up or justifies violence, hatred or intolerance,[131] or revisionist speech may not fall under the scope of protection.[132]

Although the freedom of expression has a broad scope, limitations may be acceptable if they serve to protect other weighty interests. These interests can be personal, such as reputation,[133] the right to non-discrimination,[134] or property,[135] or societal, such as discourse during elections[136] or public order.[137] In cases where those interests outweigh the interests of the expression, the right to freedom of expression may be restricted. The conditions for such restrictions to be acceptable are codified in article 10(2) ECHR and 52(1) CFR. These provisions state that restrictions must be prescribed by law, must pursue a legitimate aim of general interest, and must be necessary and proportionate to this aim.[138] The first requirement is a requirement of the rule of law; not only must restrictions be codified in law, that law must also be accessible and predictable.[139] This prevents restrictions that lack a legal basis in

[126] Andrei Marmor, 'Two Rights of Free Speech' (2018) 31 Ratio Juris 139; TM Scanlon, 'Why Not Base Free Speech on Autonomy or Democracy Responses' (2011) 97 Virginia Law Review 541. Also e.g. European Court of Human Rights, 8 July 1986, *Lingens v Austria,* 9815/82, para 41.

[127] European Court of Human Rights, *United Communist Party of Turkey and Others v Turkey,* 19392/92, para 57.

[128] European Court of Human Rights, 7 December 1976, *Handyside v the United Kingdom*, 5493/72, para 49.

[129] E.g. European Court of Human Rights, *Ahmet Yilderim v Turkey*, 18 December 2012, 3111/10; European Court of Human Rights, *Melike v Turkey*, 15 September 2021, 35786/19.

[130] European Court of Human Rights, 2 October 2012, *Rujak v Croatia,* 57942/10, 57942/10, para 29.

[131] European Court of Human Rights, 15 October 2015, *Perincek v Switzerland,* 27510/08; More broadly e.g. Natalie Alkiviadou, *Hate Speech and the European Court of Human Rights* (Routledge 2025).

[132] More extensively Janneke Gerards, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023) 86–87.

[133] European Court of Human Rights, 25 February 2022, *Biancardi v Italy*, 77419/16.

[134] European Court of Human Rights, 16 July 2009, *Willem v France,* 10883/05.

[135] European Court of Human Rights, 19 February 2013, *Neij and Sunde Kolmisoppi v Sweden,* 40397/12.

[136] European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 153; European Court of Human Rights, 16 July 2009, *Féret v Belgium,* 15615/07, para 67.

[137] European Court of Human Rights, 8 June 2010, *Gül et al v Turkey*, 4870/02, para 42.

[138] European Convention on Human Rights, article 10(2).

[139] European Court of Human Rights, 20 January 2020, *Magyar Ketfarku Kutya Part v Hungary,* 201/17, para 94. European Court of Human Rights, 26 April 1979, *Sunday Times v the United Kingdom,* para 49.

national or EU law. The second requirement is a requirement of legitimacy. Article 10(2) ECHR provides an exhaustive list of interests for which the right to freedom of expression may be limited, such as public health, or preventing disorder and crime.[140] The final condition for restriction of the right to freedom of expression is that the restriction is proportional and necessary. These principles allow the Court to weigh whether the restriction is proportional to the interest pursued and reflects a pressing social need to limit the litigious expression, or whether it is outweighed by the rights and freedoms of others.[141]

Protecting the freedom of expression is therefore a balancing act, which means that state actors must weigh societal interests and freedoms, and can be held to account if restrictions contravene the requirements of article 10(2) ECHR and article 52(1) CFR. The balancing is act is contextual, and depends on who the speaker is,[142] the topic of expression,[143] the parties affected by the expression,[144] the degree of limitation,[145] the context in which it was uttered,[146] and the potential reach of the expression.[147] The ECtHR and the CJEU have developed case law to guide state actors in how to limit expressions without unduly interfering with fundamental rights.[148]

2.2.c. Obligations for State Actors for the Protection of the Freedom of Expression

Primarily, states have a negative obligation to not interfere with the right to freedom of expression. This means not restricting the right to freedom of expression unless such interferences are in line with the requirements outlined above. However, the right to freedom of expression may also entail a positive obligation, which means that states have to ensure and facilitate an open environment in which thoughts and opinions can be exchanged.[149] Such obligations may encompass providing

[140] European Court of Human Rights, 23 October 2008, *Khuzhin et al v Russia*, 13470/02, para 117 (in the context of article 8 ECHR).

[141] European Court of Human Rights, 26 April 1979, *Sunday Times v the United Kingdom,* para 62.

[142] E.g. more freedom may be awarded to the press: European Court of Human Rights, 27 June 2017, *Case of Stakunnan Markkinapörssi Oy and Satamedia Oy v Finland,* 931/13.

[143] As stated above, some topics may be excluded from protection, such as xenophobic, Islamophobic, or homophobic expressions; if they are not, states can enjoy a wider margin of appreciation in restricting that speech, see European Court of Human Rights, 9 February 2012, *Vejdeland et al v Sweden,* 1813/07, para 57.

[144] E.g. politicians may be subject to more scrutiny than private individuals, see European Court of Human Rights, 8 July 1986, *Lingens v Austria,* 9815/82, para 42.

[145] E.g. European Court of Human Rights, 24 September 2003, *Appleby et al v the United Kingdom,* 44306/98.

[146] E.g. European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09, para 115-116.

[147] E.g. the increased reach over the internet can increase the responsibility of the speaker, see European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 193.

[148] E.g. Pitruzzella and Pollicino (n 15), 50-53.

[149] Gerards (n 132) 101. Referring to European Court of Human Rights, 14 September 2010, *Dink v Turkey,* 2668/07, para 137-138.

information on a topic of general interest to public watchdogs,[150] or ensuring media pluralism through broadcast licensing practices.[151] A positive obligation toward the protection to freedom of expression can also mean upholding the right to freedom of expression in affairs between private actors, for example through weighing such interests in case law. Examples are found in *Google Spain v Mario Gonzalez*[152] on the right to information, or *Youtube/Cyando* on the right to freedom of expression.[153] The next subsection discusses how the freedom of expression plays a role in affairs between users and internet intermediaries.

2.2.d. Obligations for Internet Intermediaries for the Protection of the Freedom of Expression

Private actors have an important position in governing and protecting the freedom of expression on the internet.[154] This creates a tension with the protection of freedom of expression as laid down in fundamental rights instruments such as the ECHR, the CFR, and national constitutions, since private entities *prima facie* do not have an obligation to protect fundamental rights. Fundamental rights instruments such as conventions, treaties, and constitutions commonly bind only state and supra-national organisations, but not private actors. This subsection addresses how private actors can still be held to norms around the freedom of expression.

The EU regulator has attributed responsibilities for fundamental rights to private actors. For example, the General Data Protection Regulation (Regulation 2016/679, 'GDPR'[155]) imposes an obligation on private actors to respect and protect the right to data protection.[156] Similarly and crucially for this dissertation, hosting services bear responsibility for Charter rights in their enforcement of their terms and conditions under article 14(4) DSA.[157] In such instances, the internet intermediary takes on a quasi-constitutional role by virtue of their powerful position in the digital realm; the regulator seeks to limit that power by attributing responsibility for fundamental

[150] European Court of Human Rights, 8 November 2016, *Magyar Helsinki Bizottsag v Hungary*, 18030/11, para 159.

[151] European Court of Human Rights, 5 April 2022, *NIT v Moldova*, 28470/12, para 190.

[152] Court of Justice of the European Union, 13 May 2014, *Google Spain v Mario Gonzalez*, C-131/32; Gerards suggests that it could be argued that freedom of information has a horizontal effect in EU law: "the obligation to take action to ensure the freedom of information or balance it against other rights is, after all, directly imposed on the search engine operator."

[153] Court of Justice of the European Union, 22 June 2021, *Peterson v Google/Elsevier v Cyando*, C-682/18 & C-683/18, para 116.

[154] Klonick, 'The New Governors' (n 18). Also European Court of Human Rights, 16 June 2015, *Delfi v Estonia*, 64569/09, para 113 on the differing role of internet intermediaries to normal news outlets.

[155] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), *OJ L 119*, 4 May 2016, p. 1-88.

[156] e.g. Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (1st edn, Cambridge University Press 2022) 249.

[157] E.g. Regulation 2022/2065, art. 14(4), see also João Pedro Quintais, Naomi Appelman and Ronan Ó Fathaigh, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' [2023] German Law Journal 1.

rights. Article 14(4) DSA can be seen as a codification of the use of open-ended norms in tort- or contractual liability has in national law to create an indirect responsibility for fundamental rights such as the freedom of expression.[158] For example, the interpretation of the private law relation between a user and a platform may be unreasonable from the perspective of the user's freedom of expression – which is now safeguarded in article 14(4) DSA.

The quasi-constitutional role of internet intermediaries is interesting because most content moderation decisions taken by internet intermediaries are based on a contractual relation with the user, namely the terms and conditions of the internet intermediary, and not on the law.[159] These are the rules that the internet intermediary unilaterally formulates and that often go beyond what is deemed formally illegal by national laws. For example, some forms of pornography might be accepted as a legitimate exercise of the freedom of expression under fundamental rights law and may not be punishable under national criminal laws, but many internet intermediaries still remove it to keep their platform accessible to minors and retain advertisements revenue.[160] Internet intermediaries enjoy the freedom to set their own rules, based on their right to property, their freedom to conduct a business,[161] and even their freedom of expression.[162] However, it means that expressions online are subject to self-regulation that may be unilaterally changed, which creates risks in itself, and these rules may not always represent an adequate balance of freedom of expression and private interests.[163] Whether these rules are adequately applied in respect of fundamental rights or form a disproportional limitation of such rights can be adjudicated and corrected only after the fact in the aforementioned interpretation of open-ended norms in tort- or contractual liability.

---

[158] E.g. Bundesverfassungsgericht, 22 May 2019, *Der III. Weg*, 1 BvQ 42/19.

[159] On the overwhelming proportion of content moderation decisions being taken on the basis of terms and conditions, see Chapter II. On the governance through contractual relations, see e.g.: Niva Elkin-Koren, Giovanni De Gregorio and Maayan Perer, 'Social Media as Contractual Networks: A Bottom Up Check on Content Moderation' (2022) 107 Iowa Law Review 987; Mateusz Grochowski, 'Shadow Contract Law in the Platform Economy' in Federica Casarosa and Mateusz Grochowski (eds), *Enforcing Private Regulation in the Platform Economy* (Mohr Siebeck 2025).

[160] Although in such cases too, tensions can arise, for example the removals of a Pulitzer prize-winning image of a Vietnamese girl fleeing a Napalm attack: 'Fury over Facebook "Napalm Girl" Censorship' (n 21) or the removal of breastfeeding footage: Yasmin Ibrahim, 'The Breastfeeding Controversy and Facebook: Politicization of Image, Privacy and Protest' (2010) 1 International Journal of E-Politics (IJEP) 16.

[161] Berdien van der Donk, 'Restricting Access to Legal Content in House Rules: The Platform's Freedom To Conduct a Business as a Counterargument To Limit Users' Freedom of Expression', *Triangulating Freedom of Speech: Business, Social Rights, and the Freedom of Speech in a Digital Age* (Springer 2021).

[162] E.g. European Court of Human Rights, 8 July 2025, *Google v Russia*, 37027/22.

[163] Compare Regulation 2022/2065, art. 14(4), see also Quintais, Appelman and Ó Fathaigh (n 157).

The DSA aims to strengthen the protection of fundamental rights in the digital realm. It constitutes a regulatory intervention at a systemic level, and creates a procedural framework for individual rights. The DSA emphasises the protection of individual fundamental rights in content moderation by continuous reference to the Charter, and stresses the need for internet intermediaries to moderate content and conduct their services with respect for the fundamental rights of internet users. It also establishes a framework for content moderation that instils aspects of the right to an effective remedy in the content moderation process. It introduces procedural rights for internet users, legal certainty with regards to terms and conditions, and the potential for independent adjudication of content moderation disputes, and seeks the protection of fundamental rights through that network.[164] Part of that procedural framework of content moderation in the DSA is a complex network of actors, establishing procedures between internet users, internet intermediaries, civil society stakeholders, and state actors. To understand this network of actors, this dissertation builds on theories from media- and platform studies that describe the network of actors and the processes governing them, how they shape content moderation, and ultimately how they shape freedom of expression. These are explored in the next subsection.

2.3. Multistakeholderism and Networked Platform Governance

Internet regulation has the tendency to be the result of interactions between a wide range of actors. This can be seen as multistakeholder governance. Farrand and Carrapico argue that "the regulation of the internet is the result of complex relationships between both public and private actors acting both in coordination and competition with each other in order to ensure preferred regulatory outcomes."[165] The DSA is no different. Several authors have noted that the DSA is a formalisation of a network of actors in the content moderation process.[166] These actors can be seen as regulatory intermediaries.[167] Regulatory intermediaries, as defined by Medzini and Levi-Faur, are actors "that possess certain regulatory responsibilities and are situated in the regulatory regime between regulators, regulated organisations, or

[164] Husovec (n 70) 33–54.

[165] Benjamin Farrand and Helena Carrapico, 'Networked Governance and the Regulation of Expression on the Internet: The Blurring of the Role of Public and Private Actors as Content Regulators' (2013) 10 Journal of Information Technology & Politics 357, 363.

[166] Jens-Peter Schneider, Kester Siegrist and Simon Oles, 'Collaborative Governance of the EU Digital Single Market Established by the Digital Services Act' (2023) 11; Dvoskin (n 29) 457; Rachel Griffin, 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability' (2023) 14 Transnational Legal Theory 46, 53; De Gregorio and Dunn (n 34) 486.

[167] Beatriz Kira, 'Regulatory Intermediaries in Content Moderation' (2025) 14 Internet Policy Review .

other groups of stakeholders such as beneficiaries."[168] These actors are formally afforded competencies through regulation, and act as an intermediary between the regulator and the target of the regulation. In the context of the DSA, this applies to internet intermediaries, trusted flaggers, out-of-court dispute settlement bodies, auditors, and other stakeholders.

A popular critique of multistakeholder governance is that such models are voluntary, steered by the internet intermediary, and therefore unlikely to realise any real impact.[169] The DSA overcomes this by providing rights and responsibilities for regulatory intermediaries. This also raises a concern. The shift of responsibility in content moderation means that potential risks for fundamental rights not only result from the private conduct of a single actor – the internet intermediary. Risks may also arise from how the regulation distributes power in content moderation over various actors, and how that regulation is subsequently enforced. This not only raises questions about who is attributed power in content moderation by this shift of responsibility, but also on how these actors interact, and how this interaction ultimately affects the balancing acts involved in content moderation. To understand the theory of multistakeholderism and how it manifests in forms of platform governance, this subsection provides an understanding of the involvement of different stakeholders in platform governance.

Platform governance is defined by Gorwa as "the set of legal, political, and economic relationships structuring interactions between users, technology companies, governments, and other key stakeholders in the platform ecosystem".[170] Platform governance is a multi-layered and multi-actor framework. It includes governance of online platforms by state, supra-national, or non-state actors; governance by online platforms of their users, through contractual relationships and enforcement of community guidelines; and governance through online platforms.[171] To describe platform governance, authors have relied on geometrical frameworks,[172] a system of

[168] Rotem Medzini and David Levi-Faur, 'Self-Governance via Intermediaries: Credibility in Three Different Modes of Governance' (2023) 25 Journal of Comparative Policy Analysis: Research and Practice 323, 324.

[169] Elettra Bietti, 'A Genealogy of Digital Platform Regulation' (2023) 7 Georgetown Law and Technology Review 1, 42–45; Griffin, 'Public and Private Power in Social Media Governance' (n 166) 74.

[170] Robert Gorwa, 'The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content' (2019) 8 Internet Policy Review 1, 2. See also Robert Gorwa, 'What Is Platform Governance?' (2019) 22 Information, Communication & Society 854

[171] Tarleton Gillespie, 'Governance of and by Platforms' in Jean Burgess, Thomas Poell and Alice Marwick, *The SAGE Handbook of Social Media* (SAGE Publications 2017) 2.

[172] Balkin (n 6); Gorwa (n 170); Oreste Pollicino, 'The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism' [2023] European Law Journal 1.

processes,[173] or as a framework of legal, contractual relationships.[174] This dissertation proposes looking at platform governance in the DSA as a multistakeholder effort involving a range of actors. To that end, it relies on Caplan's theory of 'networked platform governance', as this subspecies of multistakeholderism most closely resembles the relationships between the different actors that the DSA formalises in content moderation. This does not dismiss the validity of looking at platform governance through any of the frameworks above, but this specific contribution places more emphasis on the role of actors involved in platform governance and content moderation.

Subsection 2.3.a explains a general, overarching theory of multistakeholderism, and subsequently moves to Subsection 2.3.b. to explain Caplan's networked platform governance as a more specific species of multistakeholderism. This theory is explicitly applied to the DSA in Subsection 2.3.c.

2.3.a. Multistakeholderism

Multistakeholderism is an oft-touted term to describe modes of governance that involve input, in some way, from multiple stakeholders. Multistakeholder governance has been an important means of internet governance.[175] Raymond and DeNardis define it as "two or more classes of actors engaged in a common governance enterprise concerning issues they regard as public in nature, and characterised by polyarchic authority relations constituted by procedural rules."[176] An underlying assumption in their definition is that multistakeholder governance exists in many forms, and is not fully conceptually demarcated.[177] For example, multistakeholder governance efforts can be private arrangements between industry and civil society representatives, or co-regulatory effort between states and online platforms, potentially following consultation with civil society.

Raymond and DeNardis' definition consists of two core components. The first component is that different classes of actors engage in a governance enterprise. As

[173] Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 526; cf. Kate Klonick, 'Of Systems Thinking and Straw Men Online Speech: Response' (2022) 136 Harvard Law Review Forum 339.

[174] Elkin-Koren, Gregorio and Perer (n 159); cf. Catalina Goanta, 'The New Social Media: Contracts, Consumers, and Chaos' (2023) 108 Iowa Law Review 118.

[175] Generally e.g. Jeremy Malcolm, *Multi-Stakeholder Governance and the Internet Governance Forum* (1st paperback ed, Terminus Press 2008); Mark Raymond, 'Puncturing the Myth of the Internet as a Commons' [2013] Georgetown Journal of International Affairs 53.

[176] Mark Raymond and Laura DeNardis, 'Multistakeholderism: Anatomy of an Inchoate Global Institution' (2015) 7 International Theory 572, 573.

[177] ibid 575.

such, international treaties between states, or industry associations including only firms do not qualify as multistakeholderism, since these consist of the same actor class.[178] In internet governance, they identify four types of actors that can interact: state actors, supra-state actors, firms, and civil society. These generally represent an aggregate categorisation of actors. The second component is that procedural rules govern the relations between these actors. This is where, according to Raymond and DeNardis, multistakeholder governance systems can differ most.[179] In governance enterprises, the internal procedural rules of the governance enterprise can reflect the prioritisation of governance objectives. These procedural rules can affect internal power relationships, influence how actors communicate with each other, and ultimately determine the success of the collaboration.

To describe multistakeholder forms of governance it is necessary to understand the underlying rationale for that form of governance. The choice for multistakeholder governance can be motivated from different perspectives. One rationale is that involvement of different stakeholders ensures representation which prevents marginalisation of minority groups platform governance.[180] Griffin underscores this idea by adding an ideological note to Raymond and DeNardis' definition, defining multistakeholderism as "interventions seeking to increase civil society influence in social media governance",[181] highlighting that the implication of such interventions is that more civil society influence is both necessary and desirable. Another rationale is that multistakeholder governance can, in theory, help decentralise power in the digital realm.[182] When responsibility for developing and conducting content moderation processes is divided over multiple actors, the relative platform monopoly on power in content moderation is somewhat mitigated. This applies similarly to avoiding a state monopoly in internet governance. Some proponents of multistakeholder governance of the internet believe that multistakeholderism can be used to avoid state interference when the internet is governed by civil society organisations. A prime example of this ideological grounding is ICANN.[183] A final incentive for online platforms to embrace multistakeholder governance can be to

[178] ibid 579.

[179] DeNardis & Raymond make a distinction between four types of systems: hierarchy, heterogenous polyarchy, homogenous polyarchy, and anarchy. They argue that only the two forms of polyarchy are consistent with multistakeholderism, as these are the only forms where power is distributed amongst multiple actors in an orderly fashion.

[180] Gorwa, 'What Is Platform Governance?' (n 170).

[181] Griffin, 'Public and Private Power in Social Media Governance' (n 163) 50.

[182] DeNardis & Raymond (n 176) 580.

[183] See Raymond & Denardis (n 176) 594-596, generally Milton Mueller, *Declaring Indepence in Cyberspace: Internet Self-Governance and the End of U.S. Control of ICANN*, Cambridge: The MIT Press, 2025.

promote their own image as an ultimately benevolent actor, who is actively trying to make the internet a better place by involving stakeholders, creating the image of a democratic platform, while simultaneously downplaying their own power in content moderation.[184] This can also be seen as a conscious effort to stave off regulation.[185]

Multistakeholderism can be used as a form of internet governance across different layers. The internet can be disaggregated as follows: internet has an infrastructure layer, a standard-setting layer, a content layer, and a social layer.[186] The infrastructure layer concerns critical internet resources concerning aspects such as connectivity. The standard-setting layer, which coordinates technical standards by which computing devices can exchange information, has been a predominant focal point of multistakeholderism. Organisations such as ICANN govern domain names and protocols developed by organisations such as the Internet Engineering Task Force.[187] Less examples of multistakeholderism are available for the content and social layers, but organisations such as the TechCoalition are an example of a multistakeholder coalition that targets child abuse material on the internet.[188] The content- and social layers concern legal- and social aspects: what content is illegal and must therefore be removed, and what content is unwanted and should therefore be moderated. Traditionally these are topics on which more disagreement exists.

As mentioned above, networked platform governance can be seen as a nuancing of general theories of multistakeholderism. As a theory to describe platform governance it does not necessarily depart from what was written above but adds an extra layer that more extensively describes the motivations behind multistakeholder engagement and the interactions between the various stakeholders in platform governance. This theory is explored in the next subsection.

2.3.b Networked Platform Governance

In addition to framing modes of governance as multistakeholderism, several authors have also used the metaphor of 'networks' to describe how the internet is regulated. Of particular importance in this respect is a theory developed by Caplan which entails that networked platform governance describes how "companies [try] to

[184] Caplan (n 30).

[185] Kira (n 166) 5.

[186] Vint Cerf, Patrick (I) Ryan and Max Senges, 'Internet Governance Is Our Shared Responsibility' (2014) 10 I/S: A Journal of Law and Policy for the Information Society 1.

[187] Milton Mueller, Andreas Schmidt and Brenden Kuerbis, 'Internet Security and Networked Governance in International Relations' (2013) 15 International Studies Review 86.

[188] See https://technologycoalition.org/.

distribute the responsibility for content policy making to external stakeholders in response to concerns about their unilateral control and proposes that platforms are deploying a brokered form of network governance that presents opportunities for platform governance but also new concerns for accountability."[189]

Although networked governance conceptually resembles multistakeholderism, Caplan argues they are different. Referring to a report by Huppé, Creech, and Knoblauch,[190] Caplan distinguishes networked governance from multistakeholderism because of the differences in access to information, methods of communication, and motivation among stakeholders.[191] Caplan argues that networked platform governance more adequately captures how power relationships and information positions run across different stakeholders in platform governance. Networked governance would therefore more accurately describe platform governance than multistakeholderism, as it acknowledges the reality that platforms hold significantly more power and knowledge than other stakeholders involved in internet governance, such as users, civil society, or even states.[192] Looking at platform governance as governance networks allows for a mapping of multistakeholder activities of internet intermediaries, as well as "hel[p] us frame questions about the degree to which they actually represent a sharing of governance or another highly brokered form amounting to a minimal shift from what existed previously in terms of both operations and public perception."[193]

Caplan builds her framework for networked platform governance around Sörensen and Torfing's concept of networked governance.[194] They define networked governance as "a concept […] to refer to the growth (or recognition of) forms of governance that work to leverage fields of interdependent (though autonomous) actors, coordinate or allocate resources, make decisions, and achieve organisational goals."[195] Mapping this onto online platforms, Caplan proposes that platform governance involves a range of actors forming a governance network. This form of governance is distinct from traditional, hierarchical forms of government or public-

[189] Caplan (n 30) 3451.

[190] Gabriel A Huppé, Heather Creech and Doris Knoblauch, 'The Frontiers of Networked Governance' (International Institute for Sustainable Development 2012).

[191] Caplan (n 30) 3455. See also Dvoskin (n 29).

[192] Yik Chan Chin and Chen Changfeng, 'Internet Governance: Exploration of Power Relationship' (2017) <https://papers.ssrn.com/abstract=3107239> accessed 21 January 2025.

[193] Caplan (n 30) 3465.

[194] Eva Sørensen and Jacob Torfing, 'The Democratic Anchorage of Governance Networks' (2005) 28 Scandinavian Political Studies 195.

[195] Caplan (n 30) 3454–3455.

private neoliberal models. Online platforms engage with a network of stakeholders that aid in setting guidelines, such as codes of conduct or community guidelines, or that aid in decision-making in the content moderation process. The construction of such a network can be voluntary or compulsory. Voluntarily constructing a network of actors may be driven by platforms' wish to construct an image of being 'a democratic platform'. This functions as a marketing tactic similar to 'bluewashing', in which the platform uses constitutional language to phrase community guidelines to create a sense of legitimacy, or founds external judiciary-like institutions such as Meta's Oversight Board.[196] A compulsory network of actors may be established in regulatory instruments or as a result of codes of conduct. For example, the Code of Conduct on Countering Illegal Hate Speech Online prescribes multistakeholder participation.[197]

As a rationale for choosing for a networked form of platform governance, proponents of such models argue that interaction between stakeholders in governance networks increases diversity and adds expertise to platform governance.[198] Negotiated or deliberative governance models can help overcome hierarchical problems that are prevalent in internet intermediaries. Conversely, critics raise that forms of networked platform governance can obfuscate responsibility;[199] in the opaque landscape of platform governance, it may be impossible to assess influences of various stakeholders, or how internet intermediaries abuse power to advance their own interests or that of certain interest groups over others.

2.3.c. Networked Platform Governance, the DSA, and Accountability

The aim of this dissertation is not to achieve conceptual coherence in the complex debate on multistakeholderism, nor is it to provide a normative evaluation of its merits. The dissertation describes how the involvement of various actors in the DSA affects the balancing of the right to freedom of expression of internet users. The theory of networked platform governance is useful since the DSA's procedural framework mimics aspects of networked platform governance. The DSA's

[196] Evelyn Douek, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (2024) 37 Harvard Journal of Law & Technology; cf. Lucas Henrique Muniz Da Conceicao, 'A Constitutional Reflector? Assessing Societal and Digital Constitutionalism in Meta's Oversight Board' [2024] Global Constitutionalism 1.

[197] Code of Conduct on Countering Illegal Hate Speech Online +, accessible via https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online, see also Regulation 2022/2065, art. 45(1).

[198] Peter Bogason and Juliet A Musso, 'The Democratic Prospects of Network Governance' (2006) 36 The American Review of Public Administration 3.

[199] Gunther Teubner, '"And If I by Beelzebub Cast out Devils, …": An Essay on the Diabolics of Network Failure' (2009) 10 German Law Journal 395.

procedural framework facilitates interactions between different actor classes: state actors, internet intermediaries, and civil society. The DSA mitigates central power of platforms in content moderation by solidifying the influence of other actors. However, the power imbalance remains. Internet intermediaries still have more resources and information than other actors, which is in line with Caplan's notion in 'networked platform governance' that most of the power around content moderation is centred with the internet intermediary.

Networked platform governance in the DSA is shaped through a web of accountability relations between internet intermediaries, users, civil society, and state actors. Bovens defines accountability as "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences."[200] This maps well onto the notion of the DSA as a regulation creating relationships and procedures between actors in networked platform governance. The forum in this case is the network shaping platform governance. The result of platform governance is the answering of questions on what– and how content is moderated. Actors are under an obligation to explain and justify their conduct with regard to the freedom of expression of other participants in that forum, and can be questioned for that conduct by those participants through notice-and-action mechanisms or through public accountability on the provided transparency, which can be followed by consequences, such as liability or sanctions by oversight authorities. However, in complex governance networks, questions about accountability can be obfuscated, as it may be unclear who is accountable for what aspect of platform governance.[201]

If the DSA is observed as creating a procedural network for a polyarchic form of platform governance that seeks to foster accountability for platform practices,[202] the descriptive power of the theory of networked platform governance helps to interrogate the process around the involvement of actors in content moderation, and to analyse how that involvement affects content moderation. This is important from the perspective of accountability: which actor can be questioned and potentially sanctioned for the outcome of their conduct? Accountability is necessary to understand the balancing of freedom of expression in content moderation. The

[200] Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 European Law Journal 447, 450.

[201] Caplan (n 30) 3466.

[202] Husovec (n 58) 910.

involvement of each actor introduces new interests in content moderation, and each of these interests can be balanced against the freedom of expression of (other) internet users. The interactions and power relations between these actors determine who gets to balance those interests shaping the freedom of expression, how they are balanced, and what risks arise. These questions are reflected in the following chapters.

3. Approach, Outline, and Methodology

3.1. The Article-based Dissertation

This dissertation is article-based. The chapters are individual, self-standing contributions to the academic debate. At the time of writing, five have been published (Chapters II, III, V, VI, and VII) and one is under consideration (Chapter IV). For one article (Chapter II) I share authorship with Rishabh Kaushal, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi; for the other articles I am the sole author.

An article-based dissertation allows for timely contributions to a fast-moving debate. This is especially pertinent for a dissertation on technology regulation: at the start of this PhD project, the DSA was not yet formally been adopted; at the time of writing its conclusion, the debate focuses on possibilities for simplification, or easing-off on enforcement.[203] There are downsides to this approach: since every chapter is a self-standing scholarly article, there is unavoidably some overlap between the chapters. Next to that, once a contribution is published, its contents cannot be adjusted to reflect changes in law, technology, or society. In the case of Chapter VII, a short addition has been made to the article to reflect the latest developments. Other than that, the contents of the various articles have not altered, except for the purposes of standardising the referencing format to OSCOLA and the numbering of the headings.

3.2. Outline

The dissertation proceeds in three parts corresponding to the three classes of actors identified above, followed by concluding remarks. Each of the parts consists of two

[203] Ramsha Jahangir, 'What's Behind Europe's Push to "Simplify" Tech Regulation? | TechPolicy.Press' (*Tech Policy Press*, 24 April 2025) <https://techpolicy.press/whats-behind-europes-push-to-simplify-tech-regulation> accessed 24 July 2025; Catalina Goanta, 'How the White House Is Gaslighting the World About Europe's Digital Laws | TechPolicy.Press' (*Tech Policy Press*, 24 March 2025) <https://techpolicy.press/how-the-white-house-is-gaslighting-the-world-about-europes-digital-laws> accessed 7 April 2025; 'EU Should Simplify Its AI Act and Digital Rulebook: Danish Minister' (*euronews*, 15:34:17 +02:00) <https://www.euronews.com/next/2025/07/16/eu-should-simplify-its-ai-act-and-digital-rulebook-says-danish-minister> accessed 24 July 2025.

chapters. Each chapter reflects an actor in the DSA's framework for networked platform governance. Taken together, these chapters illustrate how the balancing of freedom of expression in content moderation is affected by the actors formalised in the DSA.

The first part of the dissertation addresses internet intermediaries. Chapter II provides an empirical analysis of the DSA transparency database, a novel instrument for transparency. This chapter's contribution to the dissertation and addressing the overarching research question is twofold. First, it serves as an empirical introduction that demonstrates what the content moderation landscape currently looks like. In doing so, it answers a range of descriptive questions as to who, why, how, when, and on the basis of what, moderates content. Second, it analyses the role played by transparency in networked platform governance. Transparency amongst actors is an important feature in networked platform governance, as it defines the information position between actors. It also affects the potential for accountability, which is shaped by the information position of other stakeholders. This dynamic is also present in the subsequent chapters.

Chapter III provides a legal doctrinal analysis of how the DSA has altered the core component of platform regulation that has dominated most discussions on freedom of expression in relation to content moderation: liability for hosting illegal content. It argues that, although the intermediary liability exemption provisions of the e-Commerce Directive are transplanted almost *verbatim* into the DSA, the enhanced framework of due diligence and accountability combined with the altered platform landscape compared to 2000 should be capable of limiting the liability exemption for hosting providers. This has clear effects for the freedom of expression on the internet: a shift in intermediary liability regimes may inspire platforms to moderate content more stringently, which could limit freedom of expression; conversely, may also allow a more adequate balancing of freedom of expression against other interests, meaning aggrieved parties may enjoy better remedies if their interests are affected by someone else's right to freedom of expression. Additionally, it is possible that the access to information is improved, which also forms a component of freedom of expression; Although increased moderation *prima facie* restricts the access to information, it is possible that when illegal information is removed, the quality of the public debate is improved, and therefore the right to access to information can be more effectively enjoyed.

The second part of this dissertation focuses on users and civil society. Chapter IV examines the notice-and-action mechanisms under article 16 DSA. It describes how

the formal avenue the DSA creates for users to influence content moderation – and therefore the balancing of freedom of expression – is shaped by internet intermediaries, which can affect its effectiveness. This contributes to the dissertation because notice-and-action marks an important interaction between users and internet intermediaries that shapes freedom of expression on the internet. The notice-and-action mechanism is a process facilitated by the internet intermediary through which users can define the scope for freedom of expression on the internet, through notifying internet intermediaries about illegal or unwanted content. The notice-and-action mechanisms also form a core component of the intermediary liability exemption discussed in Chapter III. Chapter IV adds to this by providing an empirical assessment of notice-and-action mechanisms of AliExpress, TikTok, Twitch, and X, using a walkthrough methodology. The findings in the chapter question platform compliance with article 16, but also allow for a legal doctrinal discussion about the importance of functional notice-and-action mechanisms in the framework of intermediary liability exemption.

Chapter V addresses the role of trusted flaggers in the DSA. Trusted flaggers are entities that can influence the content moderation process and the balancing of freedom of expression by virtue of the privileged position article 22 DSA affords them. Article 22 DSA establishes that both governmental- and non-governmental entities may be accredited by national supervising authorities as trusted flaggers. Notices by trusted flaggers may be removed more readily by virtue of their privileged position and their expertise. This means that the involvement of trusted flaggers in content moderation can affect how the freedom of expression and other interests are balanced. The framework of article 22 DSA raises questions on how the framework of trusted flaggers can be best operationalised to ensure that civil society actors can effectively partake in the fight against illegal content online, but also around how the selection of such actors can affect the scope of freedom of expression and the balancing exercises made in this regard. The chapter suggests safeguards that can be implemented in the selection stage and the operationalisation stage that ensure effective participation with adequate safeguards for freedom of expression.

The third part of the dissertation addresses the role of state actors in networked platform governance. Chapter VI addresses the processes by which Digital Services Coordinators enforce the DSA, how they influence content moderation, and how this affects the freedom of expression. Digital Services Coordinators form a network of national authorities that are tasked with enforcing the DSA. They are afforded significant powers that interact with content moderation that allow them to shape

and protect the freedom of expression of European internet users. However, the operationalisation of that framework raises question on potential differences of scope of freedom of expression in the European Union and different levels of enforcement across member states, which both can cause differences in the level of protection for freedom of expression users receive from DSA enforcement.

Chapter VII explores how state actors' informal participation in co-regulatory efforts affects the freedom of expression. Such co-regulatory efforts form a core part of networked platform governance. It theorises that such participation, while informal, can in fact exude pressure on internet intermediaries to change their content moderation practices. This phenomenon is labelled 'jawboning'. The chapter theorises that such informal pressures constitute an interference with the freedom of expression of social media users, which is not sufficiently accounted for in light of the requirements of article 10(2) European Convention on Human Rights. The chapter thus illuminates how informal interactions in setting codes of conduct affect the freedom of expression of internet users.

Finally, Chapter VIII concludes the dissertation. It summarises the findings of the previous chapters, and synthesises how networked platform governance in the DSA affects the balancing of freedom of expression in content moderation. Although this dissertation is not aimed at making normative recommendations, Chapter VIII also proposes some solutions that create accountability about how the actors studied in this dissertation affect the balancing of freedom of expression in content moderation.

3.3. Methodology

This dissertation addresses its research questions using a multimethod approach. The overarching research question, namely how the processes for various stakeholders in the DSA affect the right to freedom of expression of internet users, requires both a legal methodology and empirical research methods to describe what the DSA does. Hereafter, the methods and approach adopted in this dissertation are discussed. In addition, the individual chapters, where appropriate, introduce and explain the specific methodology used to write them.

The dissertation primarily relies on legal doctrinal methods, especially in Chapters III, V, VI, and VII.[204] It uses and interprets authoritative legal sources: laws, regulations, and case law. Authoritative sources are primarily derived from the

[204] Jan M Smits, 'What Is Legal Doctrine?: On The Aims and Methods of Legal-Dogmatic Research' in Rob Van Gestel, Hans-W Micklitz and Edward L Rubin (eds), *Rethinking Legal Scholarship* (1st edn, Cambridge University Press 2017); Sanne Taekema, 'Theoretical and Normative Frameworks for Legal Research: Putting Theory into Practice' [2018] Law and Method.

European Union legal order. It also interprets self-governance efforts by internet intermediaries following a similar method. As noted above, internet intermediaries have significant self-regulatory practices.[205] These are evaluated by themselves, but also in light of the regulatory framework around these practices. This reasoning also applies to co-regulatory efforts, where governance is a shared effort between internet intermediaries and regulators. The predominant research approach taken in this dissertation is descriptive, as it seeks to set out how the DSA and the actors it formalises alter the balancing of freedom of expression in content moderation. However, in most chapters, interpretations of the DSA are suggested or submitted that ensure that the stakeholders the DSA introduces are accountable for how they alter the balancing of freedom of expression, or how the interests of or represented by various stakeholders are balanced against the right to freedom of expression as laid down in the Charter and the Convention.

Because legal doctrinal research does not fully cover how the paper reality of the DSA manifests in practice, the dissertation further relies on computational and qualitative methods to describe content moderation more holistically. Chapter II uses insights from data science and computational methods to describe and interpret the DSA Transparency Database and provide evidence for the doctrinal arguments in the article. The multidisciplinary methodological approach of this chapter is the result of a collaboration between legal scholars and computer scientists. Chapter IV also uses a multimethod approach: it uses an empirical qualitative method inspired by media studies (the walkthrough method)[206] to qualitatively describe notice-and-action mechanisms implemented by various platforms to provide evidence for the legal doctrinal arguments around compliance with article 16 DSA and the effects on liability exemption of article 6 DSA.

Aside from legal sources, this dissertation draws on social media studies, media studies, and computer science. These sources provide a socio-technical perspective on platform regulation and describe how regulation manifests in practice, as well as how trends in the use of online platforms lead to regulatory choices. They therefore give an insight into the motivations for- and effects of platform regulation, which can help to describe how the roles for stakeholders in the DSA have developed in the

[205] Hannah Bloch-Wehba, 'Global Platform Governance: Private Power in the Shadow of the State' (2019) 72 SMU Law Review 28; Teresa Quintel and Carsten Ullrich, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond' in Bilyana Petkova and Tuomas Ojanen, *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020); Gorwa, 'The Platform Governance Triangle' (n 173).

[206] Ben Light, Jean Burgess and Stefanie Duguay, 'The Walkthrough Method: An Approach to the Study of Apps' (2018) 20 New Media & Society 881.

way they have. The understanding of such stakeholders can then aid in assessing how interactions by and with stakeholders affect the freedom of expression of internet users.

3.4. Contribution

The purpose of this dissertation is to explore how the inclusion of different actors in content moderation in the DSA affects the freedom of expression of internet users. As a whole, the dissertation describes how the DSA's innovative approach in regulating networked platform governance has implications for the freedom of expression of internet users. Findings indicate that advancing the interests of one actor may put the interests of another actor in the same procedural framework at risk, which may affect how the freedom of expression is shaped or enjoyed on the internet. This can be problematic as regulatory interventions in the DSA do neither always sufficiently account for how the freedom of expression is safeguarded in that process, nor acknowledge how interactions between stakeholders affect the balancing act with regards to freedom of expression, in spite of general references to the Charter. The overall contribution of this dissertation is to illustrate these blind spots by showcasing interventions by the DSA that display different angles of the same problem. The dissertation as a whole therefore functions as a starting point for a discussion on how further implementation and enforcement of the DSA can resolve this problem, and how it can be ensured that the DSA can achieve its aims of protecting a safe, predictable, and trustworthy online environment.

3.5. Limitations

This dissertation has its limitations. These start with the core question of this work, namely who determines what we can say online and how. This question avoids most, if not all, discussions on the material scope of freedom of expression, and focuses fully on the procedural framework around content moderation and how that framework and the actors within it shape the freedom of expression. Material discussions would be an interesting continuation of this research, but are currently outside of its scope.

The dissertation further focuses on specific actors: hosting providers, users, trusted flaggers, digital services coordinators, and state actors. As a result, it has not fully mapped the influences and interactions of every actor involved in content moderation through the DSA. Covering all actors in the DSA would have been unfeasible in view of the limitations of a PhD dissertation. It therefore has been opted for to concentrate on the actors that are the most relevant in the current academic debate and current affairs. Actors such as auditors, out-of-court dispute

settlement bodies, the Board of Digital Services Coordinators, the European Commission, and even internet intermediaries higher up the internet stack such as mere conduit and caching services remain undiscussed, but nevertheless deserve attention in future research.

In addition to that, in spite of the DSA's efforts to increase transparency around content moderation, much remains unknown, especially when it comes to informal interactions that affect content moderation, or the personal preferences of those that own internet intermediaries. Consequently, even an exhaustive mapping of interactions in the DSA would not fully account for the influences of various stakeholders in content moderation and how those relate to the freedom of expression.

Furthermore, the dissertation has defined rather generic actor classes and discusses them on a rather abstract level, while more granular analyses could be made within these classes. For example, further research could explore the responsibilities regarding notice-and-action mechanisms for specific platform types, as opposed to the most-different-case sample discussed in Chapter IV. 'Internet users' have been described throughout this dissertation as a generic class, but as a category it encompasses a wide range of actors that have different interests, rights, and responsibilities. Internet users can be businesses, who are limited in their freedom of expression by virtue of the commercial nature of their activity, influencers, who fall within a currently ill-defined scope for protection of freedom of expression for a hybrid of commercial and private actors,[207] or politicians, who in standing case law have a more expansive scope for freedom of expression, but also have additional responsibilities in light of the reach of their speech. Each of these roles comes with different 'strings attached'.

Aside from a more granular approach to the actors studied in this dissertation, there is a strong temporal element that could inform further research. The DSA was adopted only three years ago, and entered into force a year ago at the time of writing. Within this timeframe, the Commission has adopted a number of Delegated Acts[208] and Codes of Conduct[209] that further regulate the enforcement of the DSA,

[207] E.g. Giovanni De Gregorio and Catalina Goanta, 'The Influencer Republic: Monetizing Political Speech on Social Media' [2022] German Law Journal 204.

[208] E.g. Delegated Regulation supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines, C(2023) 6807 final; Delegated Act on Data Access under the Digital Services Act, C(2025) 4340 final.

and a number are still expected. These can all be scrutinised for how they affect the dynamic between the various actors in networked platform governance. Time will also tell how effectively the DSA can be enforced in a sensitive geopolitical climate. The DSA has been a contentious instrument from the perspective of the Trump Administration, and has been criticised as a censorship instrument to extract wealth from platform companies.[210] It has also been proposed as a bargaining chip[211] in a complex geopolitical environment that would require the EU pausing enforcement or simplifying the DSA to appease foreign administrations.[212] The implications of this ongoing discussion are relatively under-explored throughout, as they are difficult to predict.

The development of DSA enforcement also leaves questions open for the future. DSA enforcement is currently in its infant's shoes, with most DSCs figuring out their approach, and some DSCs still needing to be certified. The Commission has currently initiated proceedings against five member states for not (fully) designating their DSCs.[213] Developing and standardising procedures across member states will be a similarly challenging endeavour,[214] which needs to be followed carefully by academics. At the same time, the DSA is part of a grander digital strategy in the Digital Acquis. This means that enforcement also needs to co-exist and synergise with enforcement of the GDPR, the AI Act, the Data Act, the Digital Fairness Act, and the Digital Markets Act.[215] This can be challenging, since each of these instruments can be enforced by different national authorities, and moreover, can, because of broad phrasing, be used to tackle a wide range of issues.

---

[209] Code of Conduct on Countering Illegal Hate Speech Online+; Code of Conduct on Disinformation; Guidelines on measures to ensure a high level of privacy, safety and security for minors online.

[210] Humeyra Pamuk, 'Exclusive: Rubio Orders US Diplomats to Launch Lobbying Blitz against Europe's Tech Law' (*Reuters,* 7 August 2025) <https://www.reuters.com/sustainability/society-equity/rubio-orders-us-diplomats-launch-lobbying-blitz-against-europes-tech-law-2025-08-07/> accessed 12 August 2025.

[211] Claes De Vreese, Jakob Ohme and LK Seiling, 'Will Europe Sacrifice the Digital Services Act in Negotiations with Trump? | TechPolicy.Press' *Tech Policy Press* (25 March 2025) <https://techpolicy.press/will-europe-sacrifice-the-digital-services-act-in-negotiations-with-trump> accessed 7 August 2025.

[212] Anupriya Datta, 'Virkkunen: All Digital Rules to Be Reviewed in Simplification Push' (*Euractiv,* 27 March 2025) <https://www.euractiv.com/section/tech/news/virkkunen-all-digital-rules-to-be-reviewed-in-simplification-push/> accessed 7 April 2025; Jahangir (n 203).

[213] 'Commission decides to refer Czechia, Spain, Cyprus, Poland and Portugal to the Court of Justice of the European Union due to lack of effective implementation of the Digital Services Act, (*European Commission,* 7 May 2025) <https://digital-strategy.ec.europa.eu/en/news/commission-decides-refer-czechia-spain-cyprus-poland-and-portugal-court-justice-european-union-due> accessed 27 October 2025.

[214] See Chapter VI.

[215] Simona Demková and Giovanni De Gregorio, 'The Looming Enforcement Crisis in European Digital Policy' (*Verfassungsblog*, 10 February 2025) <https://verfassungsblog.de/the-looming-enforcement-crisis-ai-dsa-eu/> accessed 28 February 2025.

Given the technical, economic, social, and legal complexity of online platforms and their impact on markets and society, content moderation and the role of the DSA therein could undoubtedly benefit from (more) multi- and interdisciplinary research. On the enforcement side, this is already visible in the efforts of the European Commission to strengthen their enforcement teams with technical knowhow through the European Centre for Algorithmic Transparency.[216] Insights from computer science and media studies can help legal scholars make sense of how the reality of content moderation maps onto the legal reality, and often there is a discrepancy. For example, on paper, research data access under article 40 looks attractive for researchers to increase for accountability in platform governance, but in practice, there are hurdles that are not foreseen in the 'legal reality' of the articles' phrasing.[217] There is a growing number of contributions approaching the DSA from non-legal angles, and undoubtedly countless more will be needed to truly understand platform governance.

[216] 'European Centre for Algorithmic Transparency', (*European Commission*) <https://algorithmic-transparency.ec.europa.eu/index_en> accessed 27 October 2025.

[217] Catalina Goanta and others, 'The Great Data Standoff: Researchers vs. Platforms Under the Digital Services Act' (arXiv, 2 May 2025) <http://arxiv.org/abs/2505.01122> accessed 23 June 2025.

Part I: The Platform

"For instance, we know only as much as platforms tell us. This is not good enough. They have to open up and offer more evidence that the measures they have taken are working well. They also have to enable the public to identify new threats independently. We invite them now to provide monthly reports with more granular information than ever before."

Vera Jourová, 10 June 2020

Chapter II: Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database

Published as: Rishabh Kaushal, Jacob van de Kerkhof, Catalina Goanta, Gerasimos Spanakis, and Adriana Iamnitchi, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', (2024) *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1121-1132.[1]

Abstract: *The Digital Services Act (DSA) is a much awaited platforms liability reform in the European Union that was adopted on 1 November 2022 with the ambition to set a global example in terms of accountability and transparency. Among other obligations, the DSA emphasizes the need for online platforms to report on their content moderation decisions ('statements of reasons' – SoRs), which is a novel transparency mechanism we refer to as automated transparency in this study. SoRs are currently made available in the DSA Transparency Database, launched by the European Commission in September 2023. The DSA Transparency Database marks a historical achievement in platform governance, and allows investigations about the actual transparency gains, both at structure level as well as at the level of platform compliance. This study aims to understand whether the Transparency Database helps the DSA to live up to its transparency promises. We use legal and empirical arguments to show that while there are some transparency gains, compliance remains problematic, as the current database structure allows for a lot of discretion from platforms in terms of transparency practices. In our empirical study, we analyze a representative sample of the Transparency Database (131m SoRs) submitted in November 2023, to characterise and evaluate platform content moderation practices.*

[1] Catalina Goanta is supported by the ERC Starting Grant research project HUMANads (ERC-2021-StG No 101041824) and the Spinoza grant of the Dutch Research Council (NWO), awarded in 2021 to José van Dijck, Professor of Media and Digital Society at Utrecht University.

The individual contribution to work has been the conceptualisation of the legal framework and theories of transparency, and drafting of the introduction, prior work, and DSA compliance sections.

1. Introduction

The Digital Services Act[2] is a much awaited platform regulation reform in the European Union that was adopted on 1 November 2022, with the ambition to set a global example in terms of accountability and transparency. Among other obligations, the DSA emphasizes the need for online platforms to report on their content moderation activities. In particular, Article 17 and 24 DSA govern so-called 'statements of reasons' (SoRs) that are "clear and specific" explanations about content moderation decisions. Anonymized SoRs must be submitted to the Commission without undue delay, for their inclusion in "a publicly accessible machine-readable database managed by the (European) Commission" (Article 24(5) DSA). SoRs are currently made available in the DSA Transparency Database,[3] launched by the European Commission in September 2023. Very Large Online Platforms (VLOPs) already need to comply with these obligations (Article 92 DSA). These are online platforms that have at least 45 million monthly active users in the European Union. As of 17 February 2024, all online platforms in the European Union engaged in content moderation will be under an obligation to submit SoRs to the Commission.

The DSA Transparency Database marks a historical achievement in platform governance. The database is a first of its kind, providing unprecedented insights into an otherwise entirely opaque world of content moderation,[4] dominated by standards exclusively designed by platforms. This paper focuses on understanding whether the database delivers on the transparency objectives of the DSA with respect to SoRs as a novel computational transparency mechanism, which we call *automated transparency* due to the fact that it is based on the standardization and automation required in providing large-scale data and information about content moderation decisions. To emphasize the practical role of automated transparency, we design two compliance-related questions, that we subsequently investigate through a multidisciplinary methodological approach using the legal doctrinal method for legal argumentation, and data science insights as evidence for the arguments made:[5]

---

[2] European Parliament. 2022. 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). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065.

[3] https://transparency.dsa.ec.europa.eu/

[4] Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.

[5] Jan M Smits, 'What Is Legal Doctrine?: On The Aims and Methods of Legal-Dogmatic Research' in Rob Van Gestel, Hans-W Micklitz and Edward L Rubin (eds), *Rethinking Legal Scholarship* (1st edn, Cambridge University Press 2017).

RQ1: To what extent does the Transparency Database structure (hereafter called 'schema') as operationalized by the Commission, reflect the conditions of Article 17 DSA?

RQ2: What are the compliance issues of content moderation practices as reflected by VLOP submissions in the Transparency Database in relation to Articles 17 and 24(5) DSA?

In answering these questions, we make two main contributions. *First*, we investigate and analyse the database schema, to better gauge what information VLOPs are required to provide, and how this information complies with the specific requirements of Articles 17 and 24 DSA. By comparing legal requirements with metadata categories (called 'attributes' in the schema), we can identify the transparency gains but also potential limitations of the provided data, in terms of format and information depth. *Second*, we empirically describe the information submitted through SoRs using exploratory data analysis techniques. We use a data sample (131m SoRs) reflecting all VLOPs in the Database at the time of retrieval, and formulate a series of data analysis tasks which provide further evidence relating to the early DSA-compliance strategies of VLOPs in their content moderation decisions.

2. Prior work and contributions of the present study

Launched on 25 September 2023,[6] the DSA Transparency Database is a relatively untapped resource. Previous work has focused predominantly on describing trends in content moderation on social media. Drolsbach and Pröllochs[7] used 156 million SoRs to describe content moderation practices for social media platforms over two months from the lens of data analysis, by looking at the different metadata reported in the Database. Dergacheva et al.[8] use a different temporal scope in their study, analyzing 24 hours of SoRs for social media platforms (2 million SoRs). Their findings reflect those in Drolsbach and Pröllochs,[9] and provide additional analyses

[6] https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-launches-transparency-database

[7] Chiara Drolsbach and Nicolas Pröllochs, 'Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database' (arXiv, 7 December 2023) <http://arxiv.org/abs/2312.04431> accessed 14 December 2023.

[8] Daria Dergacheva and others, 'One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database' (Center for Media, Communication and Information Research 2023) <https://platform-governance.org/2023/one-day-in-content-moderation-by-social-media-platforms-in-the-eu/> accessed 14 December 2023.

[9] Drolsbach and Pröllochs (n 7).

for content categories and the nature of illegal speech. Lastly, Trujillo et al.[10] analyse 195 million SoRs submitted by social media platforms on similar parameters (excluding those featuring free text), additionally providing cross-referencing from separate transparency reports. Similar to the other two studies, Trujillo et al. focus on social media platforms exclusively. Lastly, Kosters et al.[11] performed a theoretical legal analysis of different levels of transparency obligations focusing on a single VLOP (TikTok).

Our study complements earlier works in three respects. First, previous empirical analyses have focused on SoRs from social media platforms only, as they have been the primary object of study in content moderation. Our contribution analyzes SoRs from *all* VLOPs across different categories: *shopping platforms* (Google Shopping, AliExpress, Zalando, Amazon), *social media* (TikTok, Facebook, Pinterest, YouTube, Instagram, Snapchat, X, LinkedIn), *app stores* (Google Play, App Store) and *service platforms* (Google Maps, Booking.com). Second, our contribution is multidisciplinary, combining and contextualizing computational measurements with a legal dimension that critically reflects on the role of transparency in regulation. Finally, our descriptive exploratory analysis does not only focus on the pre-set attributes in SoRs, but we also provide further insights on the basis of linguistic data analysis on free text fields of DSA Transparency Database, to better understand platform submission practices and strategies.

3. Dataset

Since its launch, the Transparency Database has been constantly updated in real time by VLOPs, who push data into the database at varying rates. Currently, all records are available for every day from the inception of the database. Because of the massive amount of data (as of January 20, the database contains 2,8 billion records), we focus on a ten-day window between 19 and 28 of November 2023. This dataset contains more than 131 million SoRs with an average of 12 million SoRs submitted each day. We note that this is an early snapshot of the database, as all online platforms in the EU (and not just VLOPs) will be required by law to start reporting in February 2024.

Unlike other studies, we focused on a smaller time window to be able to analyse

[10] Amaury Trujillo, Tiziano Fagni and Stefano Cresci, 'The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media' (arXiv, 15 December 2023) <http://arxiv.org/abs/2312.10269> accessed 15 January 2024.

[11] Lisa Kosters and Oskar J Gstrein, 'TikTok and Transparency Obligations in the EU Digital Services Act (DSA) – A Scoping Review' (2024) 27 Zeitschrift für europarechtliche Studien 110.

SORs submitted by all platforms, not only social media platforms. Our dataset comprises varying proportions (ranging between 6.2% from Booking.com to 20% from Facebook) of total SoRs submitted by each platform as of 15 December 2023. Since our focus is more on practices and database schema as they satisfy the DSA transparency objectives, this fraction of all SORs submitted so far is sufficient for our objective.

4. Statements of Reasons (SoRs) and DSA compliance

Transparency is commonly perceived to lead to good governance, based on a twofold assumption, the first being that transparency is the most direct way of achieving accountability, the second being that transparency allows regulators to better study the phenomenon they seek to regulate. Particularly in the context of the content moderation debate, this is pertinent; content moderation is notoriously opaque,[12] and calls for transparency have been common.[13] In fact, Gillespie writes that "calls for greater transparency in the critique of social media are so common as to be nearly vacant".[14] New in the pursuit of meaningful transparency in content moderation is a database containing content moderation decisions and statements of reasons of providers of online platforms. Article 24 DSA establishes transparency reporting obligations for providers of online platforms. In particular, paragraph 5 specifies that providers of online platforms must, without undue delay, submit content moderation decisions and statements of reasons to be included in a Commission-managed database. Although the DSA enters into force for the most part from 17 February 2024, it is applicable to VLOPs 4 months after they are first designated as VLOP (art. 92), meaning that currently 17 VLOPs are covered by the DSA. The Transparency Database, as proposed in Article 24(5), is revolutionary compared to other approaches such as transparency reports or audits. Although earlier efforts offered insights on an aggregate level about the scope of content moderation over periods of 6 months or more, the ins-and-outs of content moderation have been difficult to study since there was little transparency about individual decisions so far. The DSA calls the reporting of each individual content

[12] Gorwa, Binns and Katzenbach (n 4); Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790; Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (First Harvard University Press paperback edition, Harvard University Press 2016).

[13] Sarah T Roberts, 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation' [2018] First Monday <https://www.firstmonday.org/ojs/index.php/fm/article/view/8283> accessed 15 April 2025; Nicolas Suzor and others, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 1526.

[14] Tarleton Gillespie, *Custodians of the Internet* (1st edn, Yale University Press 2018).

moderation decision as SoRs. This new transparency mechanism brings with it two novel features. First, it is a live database that needs to be updated regularly with ongoing content moderation decisions, reflecting the dynamic and hypercomplex ecosystem of content moderation. Second, it can be said to reflect raw data, and additional empirical analyses are not only possible but very much encouraged and facilitated by the Commission.[15] To this extent, it can be characterised as an automated transparency mechanism that serves as another avenue for platform data access under the DSA.

4.1. SoRs requirements

SoRs submitted by the providers of online platforms have to adhere to the requirements laid down in Article 17 DSA. Paragraph (1) requires platforms to report on restrictions imposed on users who post illegal content or content that is otherwise incompatible with their terms and conditions. Paragraph 3 lists the information that should be *at a minimum* included in the SoRs:

a. *A specification of the remedy.* Remedies may include removal or demotion of content, suspension or termination of account, suspension or termination of monetary payment, as well as the geographical and time parameters of these decisions.

b. *Facts and circumstances*. Information about whether the decision was made based on a notification under Article 16 (including the identity of the notifier where strictly necessary), or on own-initiative investigations.

c. *Automated means of decision-making and detection*. This reflects whether the platform used automated means in detecting the content or making the content moderation decision.

d. *Legal grounds*. Where the content is deemed illegal, a reference needs to be made about the legal ground used to assess the content, and why the information is considered illegal.

e. *Incompatibility with ToS*. Where the content is deemed to be incompatible with ToS, a reference needs to be made to the contractual ground and their interpretation.

f. *Redress information*. This reflects the need to include clear and user-friendly information on redress possibilities to contest the decision, such as

[15] https://transparency.dsa.ec.europa.eu

internal complaint-handling and out-of-court dispute settlement.

Paragraph 4 establishes that the information in SoRs needs to be clear and easy to understand, but also precise and specific. SoRs must be uploaded without undue delay (Article 24(5) DSA), and must be free of any personal data. Failure to submit SoRs to the Transparency Database can be interpreted as non-compliance with the DSA, opening platforms to investigations and potential sanctions by national digital services coordinators (DSCs) from the Member State where the platforms are established (Article 51(2)), as well as from the Commission in the case of VLOPs (Article 73).

4.2. SoRs template and its compliance with the DSA

To implement the requirements in Article 17, the Commission has provided a template for SoRs which it expects platforms to submit via its designated API. This format operationalizes the conditions of Article 17 as 'attributes', which can take two forms: *free textual* (with character limitations sometimes applicable), and *limited* (using pre-defined values from an array). The documentation also labels some of the attributes as mandatory three categories of attributes depending on whether they are mandatory, mandatory under certain conditions, or optional. The full SoRs schema, including the additional explanation for statement attributes is available in Section 7. In what follows, we inspect the API documentation[16] and make a general overview of the standardized attributes requested by the Commission. In January 2024, the API documentation listed a total of 33 content-related attributes. Out of these, 23 feature limited input that needs to be selected from an array, and 10 feature free textual form. In the latter category, there are seven attributes limited to 500 characters (e.g. Legal Grounds Illegal Content), two limited to 2000 characters (e.g. Illegal Content Explanation), and one limited to 5000 characters (Facts and Circumstances Relied on in Taking the Decision). In turn, out of the total 33 attributes, eight are generally mandatory, 11 are mandatory under certain logical conditions relating to other categories, and eight are completely optional. In addition, for six attributes, the API documentation is unclear on mandatory nature.

This brief investigation into the schema of SoRs gives us several important insights. First, it reveals whether the SoR standard as developed by the Commission is compliant with Article 17(3). This seems to be largely the case, with one notable

[16] https://transparency.dsa.ec.europa.eu/page/api-documentation

exception: no reference is made to redress information, in spite of the fact that it is a DSA requirement (Article 17(3)(f) DSA). Second, it reveals a balanced structure of the information SoRs are expected to deliver into the Transparency Database. However, the structure also raises further questions and concerns. Most of the fields are either conditionally mandatory, optional or unclear, which may be interpreted as a relatively high level of discretion from submitting platforms in terms of attribute choice. In addition, platforms can submit information in free textual form, and while this could potentially provide further information into platform-specific approaches, the only free textual attribute that is mandatory is 'Facts and Circumstances Relied on in Taking the Decision'. The analysis in Section 5.C offers further insights into the practices and strategies undertaken by online platforms in submitting SoRs along the attributes discussed above.

5. Platform practices and early DSA compliance

In this section, we report on the practices of the online platforms that have already submitted SoRs in the Transparency Database, and subsequently discuss the compliance implications of platform strategies in the early days of the database. Empirically, we offer evidence on multiple levels, roughly answering questions of who, how, why, what, when and where content moderation activity takes place. Specifically, we analyze *who* reporting platforms/industries are (Section 5.1); *how* their moderation practices are performed, specifically whether in an automated way (Section 5.2); *why* content is subjected to moderation actions (Section 5.3); *what* the moderation actions taken are as reported to the database (Section 5.4); *where* the content moderated comes from (in terms of language and legal territory) (Section 5.5); and finally, *when* the moderation takes place compared to the time of content posting and the time of reporting to the database (Section 5.6).

5.1. Who is Reporting: VLOPs

We investigate which platforms submit more data, and what the submission rate per platform is across the dataset. Figure 1 shows that Google Shopping has submitted more than half of the total SoRs (52.2%), followed by TikTok (17.1%), Amazon (10.8%), Facebook (9.8%), Pinterest (3.8%), Google Maps (2.6%), YouTube (1.8%), Instagram (0.7%), AliExpress (0.5%), Google Play (0.2%), Snapchat (0.08%), X (0.03%), App Store (0.005%), Booking.com (0.004%), and LinkedIn (0.003%). Given that content moderation happens across all platforms, one would expect SoRs to be approximately correlated to the number of monthly active users on social media platforms in Europe. This does not appear to be the case. SoRs submission is dominated by Google Shopping,

TikTok and Amazon, whereas these are not the largest VLOPs as reported to the Commission.[17] This opens up many questions relating to the kind of content moderation activity done within these platforms and reported in the database. It also reflects the underestimated importance of content moderation on shopping platforms.

Compliance observations. The Transparency Database provides an overview of platform practices in content moderation. However, it is difficult to estimate whether a low number of SoRs pushed in the Database can immediately translate to non-compliance. Still, in some cases, low SoR reports can be interpreted as a signal of questionable compliance in the context of additional transparency obligations.[18]

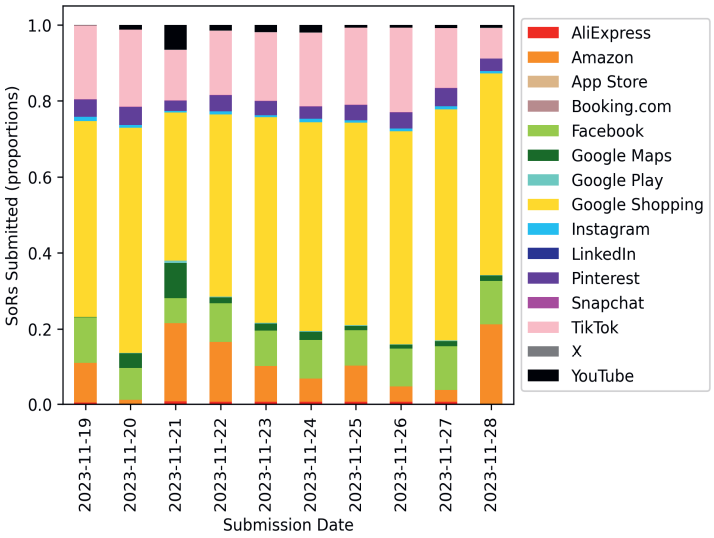


Figure 1 Fraction of SORs per day submitted by each VLOP.

[17] https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses.

[18] See for instance the Commission's legal proceedings from December 2023 against X: https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

5.2. How Moderation Decisions and Detection are Done

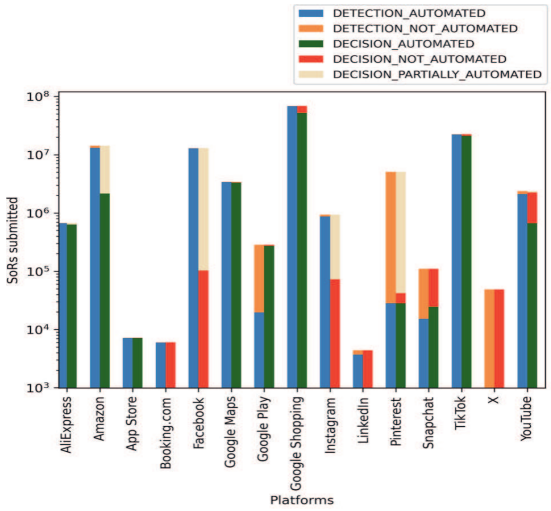


Figure 2 Distribution of automated_detection and automated_decision in SoRs

Two attributes in the SoRs describe the manner of moderation: 'automated detection' and 'automated decision'. The 'automated detection' field is binary and indicates whether automated tools were used to detect non-compliant content. Once content is detected as non-compliant, the decision on how to respond can be done in fully automated, partially automated, or in non-automated manner, which is captured in 'automated decision' attribute in SoRs. Figure 2 presents the distribution of automated detection and decision in our dataset. We find that X never reports automated means for detection, some platforms (such as Pinterest and Google Play) do not use automated means for detection in half of their SoRs, and others (such as App Store, Google Shopping and TikTok) rely mostly on automated detection. Regarding the decision for content moderation, X, LinkedIn and Booking.com do not automate at all, Facebook, Instagram, Snapchat, and YouTube make automated decisions for some SoRs, whereas AliExpress, Google Maps, Google Shopping, and TikTok make fully automated decisions for most (or all) SoRs. Also, we find that automated detection and fully automated decision are highly correlated in the moderation process, as evident in Figure 2, where the height of bars corresponding to automated detection (blue) and automated decision (green) appear correlated. When platforms do not perform automated detection (orange), then they are more likely to perform partial automated decision rather than taking no automated decision or fully automated decision.

Compliance observations. The DSA does not oblige platforms to moderate content

automatically, in contrast to other legislation.[19] However, it is generally considered desirable in the fight against harmful and illegal content that platforms undertake automated content moderation to cope with the immense volume of content shared online. This is visible in detection practices, which are predominantly automated across platforms, and partially in decision-making practices, which also involve a human touch. The involvement of humans is required by the DSA in cases of complaint procedures. Outliers who rely predominantly on human moderation are LinkedIn, Pinterest and X. The latter is currently under investigation for their content moderation practices, in which the lack of automated means can be used as a contraindication for proper moderation.

5.3. Why: Reasons for Moderation

Platforms may ground their content moderation decisions either on the law (illegal content) or their Terms of Service (ToS violations). Our dataset reveals that 99.8% of the SoRs are reported as ToS violation and only 0.2% as illegal content. *Illegal content.* For illegal content, two attributes ('illegal content legal ground', that includes up to 500 characters and 'illegal content explanation', up to 2000 characters) are specified in the Transparency Database for platforms to provide details of illegal content. Table 1 gives the distribution of the most prominent grounds ('illegal content legal ground') for marking content as illegal: AliExpress (66.2%) tops the list, followed by X (18.7%), Google Maps (8.3%), YouTube (4.3%), TikTok (2.0%), Google Play (1.1%), Facebook (0.07%), and Instagram (0.02%). Only 5 SoRs submitted by Google Shopping are flagged as illegal content. It appears that some VLOPs like Google Maps have mostly defamation related content whereas other VLOPs like TikTok have wide varying types of illegal content. We also notice that some VLOPs like Facebook specify only one standard reason[20] for reporting all illegal content. We also find that the optional attribute 'decision ground reference url' which is supposed to provide URL to the ToS or law, contains NULL values 99.9% of the times. Having said that, there are additional attributes in the Transparency Database which give more details about the illegal and ToS violation content. Further, we looked at values of 'illegal content explanation' which gives VLOPs more opportunity to give detailed explanation of grounds of flagging illegal content. To our surprise, only AliExpress makes use of this option. Out of 261,458 SoRs marked as illegal content, AliExpress gives specific and detailed explanations for 173,292 (66.2%) of them, mostly in

[19] European Parliament. 2019. Directive 2019/790 on Copyright in the Digital Single Market. https://eur-lex.europa.eu/eli/dir/2019/790/oj.

[20] For example 'This was a violation of sections 3.2 of our terms of service'.

Chinese, followed by Spanish and French. Besides AliExpress, YouTube (4%) also provides explanations with userID and product name anonymized.

ToS violations. Google Shopping (52.2%) submits more than half of the ToS violations followed by TikTok (17.1%), Amazon (10.8%), Facebook (9.8%), Pinterest (3.8%), Google Maps (2.6%), YouTube (1.8%), Instagram (0.7%), AliExpress (0.3%), and Google Play (0.2%). The trend is completely different in terms of illegal content submissions. For ToS violation content, attributes namely 'incompatible content ground' (up to 500 characters) and 'incompatible content explanation' (up to 2000 characters) are specified in the Database for VLOPs to use to provide details of ToS violation content. Table 2 gives distribution of most prominent grounds ('incompatible content ground') for marking a ToS violation content by most active VLOPs submitting ToS violation content. As an illustration, Google Shopping does not provide any URLs to their ToS for the violations. However, they provide grounds for ToS violations, most of which is due to promotional overlays on images. Likewise, TikTok mentions harassment and bullying as the main grounds for ToS violations.

Compliance observations. The trend in the grounds for the content moderation decision is that such decisions are taken for violation of the platform ToS (99,8% percent). This shows how platforms prefer to moderate on their community guidelines, rather than state law. Such a preference is not in itself non-compliant with the DSA, but it does create a paradox. The DSA places an emphasis on making what is illegal offline illegal online (recital 12). However, platforms do not moderate on what is considered 'illegal', and therefore it can be argued the DSA does not bring too much added value to the existing moderation landscape.

In addition, the 'decision facts' attribute is mandatory for SoRs, and it is central given that it describes the facts and circumstances relied on in taking the content moderation decision, which can help provide information about how platforms contextualize the complex rules they are bound to apply at scale. However, we find that this attribute mostly features generic templates that may not provide very meaningful information. Google Shopping, the most active VLOPs, submits only a standard template[21] in most (99.5%) of the SoRs. The next key contributor of SoRs, TikTok, uses a albeit shorter statement[22] in 98.5% of its SoRs. AliExpress seems to be

[21] When reviewing content or accounts to determine whether they are illegal or violate our policies, we take various information into consideration when making a decision, including product data, website quality, merchant information, account information (e.g., past history of policy violations), and other information provided through reporting mechanisms (where applicable) and own-initiative reviews.

[22] The decision was taken pursuant to own initiative investigations.

the only platform providing more specific details around moderation.[23]

Compliance observations. Article 17(1) requires SoRs to be sufficiently clear and specific. This allows users to be informed about the restrictions imposed on their content or accounts, to seek redress, and potentially alter their behaviour. SoRs can include up to 5000 characters describing facts and circumstances, and yet most such sections seem to be automatically generated and overly short. This does not necessarily allow users to identify what elements of their content are violating norms leading to restriction, nor does it allow them to form a well-argued case for redress, which remains a large problem for content moderation. Our analysis raises the specific question of whether generically formulated facts and circumstances comply with requirements of clarity and specificity. Since this is a novel question for compliance, additional research is needed to offer a more concrete answer.

VLOP	**Illegal Content (Legal Grounds)**
AliExpress	Illegal Content (67%), Knowledge Platform Governance-General Infringement (30%), Forbidden sale-prohibited ban on sale (1.6%)
X	NsfwAdultContent (47%), NsfwViolenceGore (25%)
Google Maps	Defamation (99.9%)
YouTube	Copyright removal request (92.5%), Privacy complaint (4%)
TikTok	Financial crime (47%), Hate speech (13%), Other illegal content (9%), Privacy violations (7%), Non-consensual sharing of private or intimate images (6%), Harassment or threats (3.7%), Terrorist offences/content (2.9%), Child sexual exploitation (2.3%), Defamation (2%) Content relating to violent or organized crime (1%)

Table 1: Distribution of grounds (illegal_content_legal_ground) for top-5 VLOPs relying on illegal content.

VLOP	**Ground of ToS Violation (Private Governance)**
Google Shopping	Promotional overlay on image (37%), Personalized advertising: Sexual interests (20.8%), Personalized advertising: personal hardships (12%), Restricted adult content (7.4%), Adult-oriented content (3.6%), Unsafe

[23] For instance, 'The brand name of your product [id:abc] has been selected incorrectly and therefore cannot be published'; or 'To xyz: Your product [id: abc] is similar to the def logo and important product elements, which can easily cause confusion among consumers, so it cannot be published'.

	imagery (3.2%).
TikTok	Harassment and Bullying (32.9%), Community Guidelines (26.1%), Hate Speech and Hateful Behaviors (8%), Sexual Exploitation and Gender-Based Violence (5.4%), Dangerous Activities and Challenges (4.5%), Alcohol, Tobacco, and Drugs (4.4%), Sexually Suggestive Content (2.8%).
Amazon	Violation of section S1.1 of the Amazon Business Solutions Agreement (72.7%), Violation of Amazons Listing Policies (9.5%), Violation of the German Packaging Law (VerpackG) (4.7%), Violation of the Electronical and Electronic Equipment Act (ElektroG – the German WEEE law) (3.1%).
Facebook	Violation of section 3.2 of our terms of service (100%)
Pinterest	Adult content (87.4%), Hateful activities (4.8%), Graphic violence (2.7%), Self-injury and harmful behavior (1.9%) and Medical Misinformation (1.6%)

Table 2: Distribution of grounds (incompatible_content_ground) for top-5 VLOPs relying on ToS violations.

5.4. How Platforms Sanction: What Moderation Actions are Reported

Content moderation has long been considered in binary terms, namely with content either staying up or being taken down. However, content moderation reflects a wide array of ways in which platforms can sanction undesirable content and user behavior. Most of the restrictions are done at the level of content visibility (93.4%), followed by account (6.5%), service provision (1.1%), and monetary limitations (0.015%). Every SoR has at least one of these four types of restriction. In terms of subcategories, *restriction by visibility* is further distributed among values of disabled content (66%), removed content (24%), and other (8%). Very rarely, content has been demoted (0.08%) and age restricted (0.03%). Content is mainly disabled by Google Shopping (84%), followed by Amazon (14%). Most of the content removal is done by TikTok (56%), Facebook (15%), Google Maps (12%), Amazon (8%), YouTube (7%), and Pinterest (4%). Demotion of content is only being done Facebook (98.2%) and Instagram (1.8%) whereas age restricted of content is only being done by TikTok (90.3%) and YouTube (9.7%). The most common reasons for 'others' are 'Video not eligible for recommendation in the For You feed' (50.4%), 'Limited distribution' (38.14%), 'LIVE not eligible for recommendation and restricted in search results for 10 minutes' (7.9%), and 'LIVE not eligible for recommendation and restricted in search results (1.9%). For *account restriction*, 89.9% of the times the account is suspended, and termination accounts for 10% of the times. Most of the account suspension is done by Facebook (88.9%) followed by Google Shopping (4%), Snapchat

(2%) and Amazon (2%). Most of the account termination is done by Instagram (6.2%) and TikTok (3%). Most of the *service restriction* result in partial suspension (84%), total termination (8.7%), total suspension (6%), and partial termination (0.3%). Most of the partial suspension is being done by Amazon (76%) and TikTok (7%); total suspension by Amazon (5.7%); and total termination by Google Maps (5.5%). All of the *monetary restriction* results into suspension, and all of this is contributed by Amazon. No other VLOP is performing restrictions at monetary level.

Compliance observations. The DSA does not pose strict requirements as to the type of restriction that needs to be imposed when content violates terms of service or state law. However, recital 77 stresses that, for VLOPs, unnecessary restrictions on the use of their service need to be avoided, in light of the potential fundamental rights risks that such restrictions pose. Platforms may therefore be inclined to seek restrictions that are proportional to the violation, and prefer less impactful restrictions such as demotion or demonetization as opposed to blocking account or content indefinitely where possible. This is not necessarily reflected in the dataset, but it is difficult to argue that such would be non-compliant with the DSA. Fundamental rights considerations are considerably more important for social media platforms than for other categories, such as marketplaces or app stores. From this perspective, sanctions might have a different impact. Suspending the monetization of a seller account on Amazon will be considerably different than suspending the monetization of a Live stream on TikTok. These implications need further elaboration in future research. Yet one observation is striking. Literature on platform governance has increasingly emphasized demonetization as a content moderation wrong by social media platforms.[24] Our dataset does not show any such demonetization taking place in the EU. Recital 55 refers to monetisation as the advertising revenue-driven provision of information, but the true scope of this concept remains unclear. This may be interpreted as social media platforms not yet reporting on their full monetization spectrum.

5.5. Where Content is Moderated: Language and Territorial Scope

Language is often a proxy indicator of the region from where the content was posted, or of a platform's operational language in a given geographical area. In this section, we offer some language-based insights for SoRs, and we also revisit some earlier observations in the light of these insights. The optional attribute 'content

[24] Robyn Caplan and Tarleton Gillespie, 'Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy' (2020) 6 Social Media + Society 205630512093663.

language' in the SoR schema is only populated in approximately 50% of the records in our dataset. German (DE, 34.2%) is the most frequent language, followed by English (EN, 26.5%), French (FR, 9.9%), Italian (IT, 9.3%), Spanish (SP, 3.3%), Polish (PL, 2.9%), Dutch (NL, 2.7%), Czech (CS, 2%), Romanian (RO, 1.7%), and Slovak (SK, 1.4$). Only three VLOPs, namely, Google Shopping, Google Play, and Booking.com, fill the language of content.

Additionally, SoRs feature several text fields that can be used as a more accurate proxy of the language of the moderated content. Overall, we inspect the language used in the following fields that require platforms to enter text in free form (with some limit in the number of characters): (a) *Facts and circumstances relied on in taking the decision* (decision_facts) (limit of 2000 characters, (b) *Illegal Content Legal Grounds* and *Illegal Content Explanation* are two fields that are required in a SoR if 'decision ground' was illegal content, (c) *Incompatible Content Grounds* and *Incompatible Content Explanation* are required if 'decision ground' was incompatible content. For the decision_facts, all platforms with the exception of AliExpress (where the dominant language is Chinese) report content in English. Notably, there is a difference in the length of the description with the shortest facts being provided by Instagram and Facebook (51 and 55 characters respectively) and X (59 characters) and the longest facts being provided by Google Maps/Play/Shopping (438/468/433 characters respectively) and YouTube (635 characters). These numbers also correlate with the number of unique entries by platforms in this field, e.g. Instagram and Facebook provide only 2 entries here, namely "This content was determined to violate terms of use" and "This account was determined to violate terms of use". For the Incompatible Content Grounds, all platforms except AliExpress and TikTok post content only in English (in total 13,0861,024 SoRs). For AliExpress the dominant language is Chinese (>99% of cases), while TikTok provides grounds in 31 languages. More specifically, in the cases of booking.com, Facebook, and Instagram there is only one standard entry for all relevant SoRs: Violation of article 1F in section How we work in Terms and Condition, This was a violation of "sections 3.2" of our terms of service and This was a violation of "How You Can't Use Instagram section" respectively. Notably, for the Incompatible Content Explanation and AliExpress, there is some incompatibility between the language of the grounds (Chinese) and of the explanation (could be Spanish).

For the Illegal Content Grounds 10 platforms have logged 261,458 SoRs (AliExpress, Facebook, Google Maps, Google Play, Google Shopping, Instagram, Snapchat, TikTok, X, YouTube). All platforms post in English except AliExpress where Chinese

is the dominant language in almost all entries. Facebook, Instagram, Snapchat provided only one entry for their relevant SoRs, namely "This was a violation of "sections 3.2" of our terms of service", "This was a violation of "How You Can't Use Instagram section" of our terms of use" and "Illegal or regulated activities" respectively. Other platforms like YouTube provide more extensive information (e.g. Hate Speech Law – FR or Violent Extremism Law – GR). For the Illegal Content Explanation, AliExpress has similar language patterns as with the Incompatible Content Explanation. All other platforms use English (except for very few TikTok examples TikTok which are in French).

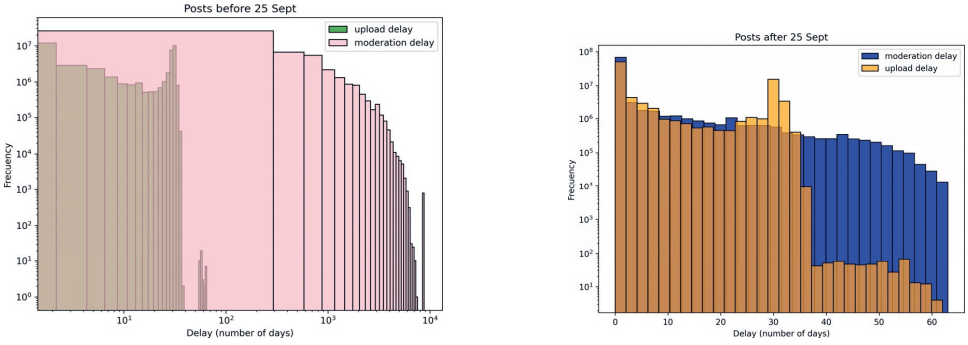
A different location-based attribute in the SoRs is 'territorial scope', which records the territorial scope of the restrictions imposed – in other words, the geographical limitations applied to content visibility, which we could also refer as the geo-blocking of content as a moderation sanction. 72.68% of the SoRs apply territorial limitations to the entire European Economic Area (EEA), followed by 16.34% to the entire European Union. There are also a few specific national level restrictions applied to Germany, France, Spain, and Italy, respectively.

Compliance observations. The DSA is silent about any language requirements relating to the submission of SoRs. In principle, it can be argued that SoRs could be submitted in any of the 24 official languages of the European Union. Following this perspective, submitting SoRs in Chinese could pose compliance questions. However, more important questions emerge when gauging the scope of the language and territoriality dimensions of the Transparency Database. The whole rationale of the Database is to provide a new, standardized tool for content moderation transparency. As we discovered, some SoRs are even submitted in multiple languages, and overall, the languages used vary significantly. This poses considerable computational constraints for the further verification of the submitted information, and future research should also be geared towards unveiling what that means in practice for the computational burden placed by platforms on public authorities and other parties interested in the DSA Transparency Database.

5.6. When SoRs are Submitted ("Without Undue Delay")

The DSA requires SoRs to be submitted without undue delay to the transparency database. 'Without Undue Delay' is a common phrasing in the DSA for a variety of platform obligations, but is not specified in this particular context. To understand the delays incurred in content moderation, we propose two metrics: moderation delay and upload delay. We define *moderation delay* as the delay between when the content was originally posted on platform and the time when content was

moderated. We define *upload delay* as the interval between the time when the content was moderated by the platform and the time when its corresponding SoR was uploaded in the Transparency Database. Note that all three time records needed for this are reported in a SoR.



(a) before 25 Sept

(b) after 25 Sept

Figure 3: Moderation and upload delay for content posted before and after the launch of the Transparency database.

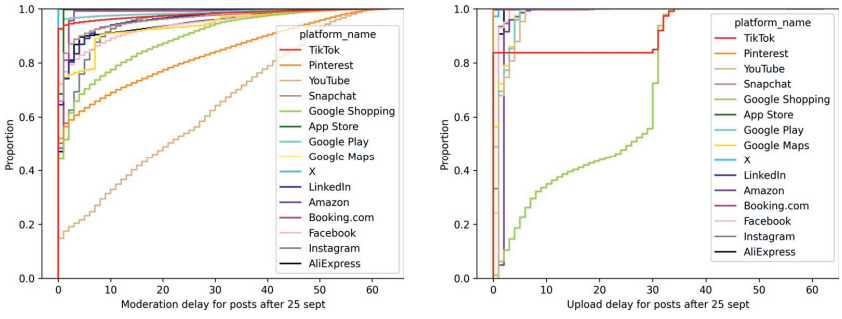
In theory, the DSA became applicable for VLOPs four months after the Commission notified the platforms it established to have VLOP designation (Article 33(6) DSA). The Commission made this designation public on 25 April 2023.[25] According to this timeline, VLOPs should have already submitted SoRs as of end of August. However, the Transparency Database was only launched on 25 September 2023, so we will consider this as the actual moment from when compliance can be considered. Once the database was launched, VLOPs submitted SoRs that predate the starting of the DSA. Overall, from the SoRs submitted on content that was posted prior to 25 September 2023, the mean moderation delay is 457.9 days (with a standard deviation of 599.6 days) while for content posted after, the mean moderation delay as reported is only 3.9 days (standard deviation 9.5 days). This significant speeding up in monitoring practices may be due to a variety of reasons, from maturing moderation practices within the company to preparing for complying with the DSA. The upload delay for content posted after the launch of the database is on average 9.6 days (with a standard deviation of 12.7 days). Details on the frequency of such delays as present in our dataset are shown in Figure 3. It appears that after the launch of the database, most of the content is moderated within a few days of posting.

Figure 4 presents cumulative distribution functions on the moderation and upload

[25] https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413

delay per platform for content posted after the launch of the transparency database. We note that Google Shopping is slower in reporting its SoRs than it is in moderating its content, which might be due to the enormous amount of reporting it has to do. (Google Shopping reports the majority of the SoRs in the database, possibly stressing the infrastructure).

Compliance observations. Article 6 DSA requires platforms to remove illegal 'expeditiously' upon learning of its existence in order to avoid liability for that content. What timeframe is meant by 'expeditiously' is undefined, however. Some practices revolved around deadlines such as 24 hours or a week. That said, this is highly dependent on the nature of the content and the circumstances of the case. It is therefore difficult to ascertain whether there is non-compliance. What is visible is that platforms that detect illegal content manually, such as Pinterest, experience *moderation delay*. Platforms that detect illegal automatically are less likely to experience *moderation delay*, with the exception of YouTube. However, it is dependent on the nature of the content restricted whether this is non-compliant with the DSA. Next to that, Article 24(5) requires platforms to submit SoRs to the database without undue delay. It is presumed that the period for submission starts when the SoR is submitted to the recipient of the service. Therefore, platforms need to avoid *upload delay*. The delay varies per platform, with the upload delay being dependent on whether the platform automates that upload or does so manually, or in batches. Compliance is ensured by automating the submission process, leading to automated transparency. Since there are no norms defined in the DSA, it is difficult to ascertain whether there is non-compliance.



(a) Moderation delay (b) Upload delay

Figure 4. Cumulative distribution function of the moderation delay (time between content posting and moderation decision) and upload delay (time between moderation decision and reporting to the Transparency Database) of content posted after the database launch

6. Conclusion: Does the DSA's automated transparency rise to its promise?

Our paper focused on investigating the DSA through the lens of a novel transparency mechanism aimed to offer insights into the opaque world of content moderation. [26] We referred to the DSA Transparency Database as a form of *automated transparency*, because it is based on a standardized process through which platforms are supposed to share structured data about their content moderation decisions. The Transparency Database is an ambitious project with no precedent. A certain degree of voluntary transparency has been commonplace for content moderation since around 2010, when Google started disclosing information on its Government Requests,[27] followed by Twitter in 2012.[28] Presently, most social media platforms endorse transparency in their content moderation practices, signing self-regulatory instruments requiring transparency such as the Santa Clara Principles on Transparency and Accountability in Content Moderation,[29] as well as co-regulatory instruments such as the EU Code of Conduct against Illegal Hate Speech[30] and the EU Code of Practice against Disinformation.[31] Regulators have also taken to the idea of transparency reporting, resulting e.g. in bi-annual transparency reporting under Germany's stringent platform liability regime.[32] Yet transparency should play a meaningful role in establishing accountability for the purpose of user benefits.[33] Transparency can be relatively meaningless if mainly used to control public discourse.[34]

Overall, we find the DSA Transparency Database to be a fascinating resource of data and information on content moderation, which aligns with earlier scholarly calls of

---

[26] Gorwa, Binns and Katzenbach (n 4); Sarah Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 New Media & Society 4366; Roberts (n 13).

[27] https://transparencyreport.google.com/government-removals/overview?hl=en.

[28] E.g. https://transparency.twitter.com/en/reports/removal-requests.html#2021-jul-dec.

[29] Prerna Juneja, Deepika Rama Subramanian and Tanushree Mitra, 'Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices' (2020) 4 Proc. ACM Hum.-Comput. Interact. 17:1.

[30] European Commission. 2019. The EU Code of conduct on countering illegal hate speech online. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and- xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

[31] European Commission. 2022. The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

[32] Amelie Pia Heldt, 'Bypassing Classical Concepts of Censorship?' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 56.

[33] Archon Fung, Mary Graham and David Weil, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007); Shagun Jhaver, Amy Bruckman and Eric Gilbert, 'Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit' (2019) 3 Proc. ACM Hum.-Comput. Interact. 150:1; Suzor and others (n 13).

[34] Mikkel Flyverbom, 'Transparency: Mediation and the Management of Visibilities' (2016) 10 International Journal of Communication 110; Ben Popper, 'Airbnb's Worst Problems Are Confirmed by Its Own Data' (*The Verge*, 4 December 2015) <https://www.theverge.com/2015/12/4/9849242/airbnb-data-new-york-affordable-housing-illegal-hotels> accessed 16 April 2025.

automating enforcement in the digital sector.[35] While by no means a perfect tool, the Transparency Database has helped reveal the sheer scale of content moderation activities, and is providing public data relating to the content moderation process in ways that so far were not accessible for civil society, researchers or public administration. In addition, the Transparency Database emphasizes the need to study content moderation from a broader perspective. So far, the content moderation debate has mostly focused on social media platforms, but the Database shows how important content moderation is for other consumer markets such as app stores, service platforms and marketplaces. Going further, content moderation studies must acknowledge this broad nature, particularly in the light of the DSA's generous definition of illegal content. By looking at the compliance of various features of the DSA Transparency Database with the DSA itself, we set out to understand how automated transparency looks like in practice in order to critically reflect on its usefulness in scrutinizing platforms and making better policy. We conclude our study with three discussion points.

Firstly, it is important to realise that the Database only provides insight in SoRs submitted by platforms, and therefore content moderation can currently only be studied to the extent that platforms allow it to be studied. In their work, Trujillo et al.[36] already found inconsistencies between the Transparency Database and the DSA-mandated Transparency Reports, showing the importance of having multiple transparency mechanisms. This has been the case earlier with various disclosure combinations such as court-mandated data sharing by platforms and voluntary transparency reports.[37] Time and time again, it has been shown that over-relying on platform power for data access leads to data manipulation risks that may defeat the entire purpose of transparently sharing the data in the first place.[38] Apart from complementary transparency mechanisms, the investigation and enforcement powers of the European Commission can also play a role in ensuring further accountability. Article 40 of the DSA, often referred to in the context of access to data

[35] Catalina Goanta, Thales Bertaglia and Adriana Iamnitchi, 'The Case for a Legal Compliance API for the Enforcement of the EU's Digital Services Act on Social Media Platforms' 9.

[36] Trujillo, Fagni and Cresci (n 10).

[37] Popper (n 34).

[38] Kim Lyons, 'Facebook Reportedly Provided Inaccurate Data to Misinformation Researchers' (*The Verge*, 11 September 2021) <https://www.theverge.com/2021/9/11/22668396/facebook-provided-inaccurate-data-misinformation-researchers> accessed 16 April 2025.

by vetted researchers,[39] is also the locus of data access powers by public authorities tasked with the enforcement of the DSA. The interplay between all possible platform insights will require further analysis and discussion. A key take-away from empirically studying the DSA Transparency Database is that platforms can strategize their use of the database as a means to show their compliance, even though in practice it may be insufficient.

Secondly, the Database creates the possibility to remain opaque on the grounds behind content moderation decisions. The DSA has attempted to make 'what is illegal offline, illegal online', reflected in recital 12 and Article 3(h). This is a laudable standpoint from a legal certainty perspective aimed to align private governance with legal standards. However, as we have seen in our analysis, most content moderation decisions in the Transparency Database (>99,8%) are based on ToS infringements. On the one hand, this can be interpreted as platforms shirking away from any responsibility of interpreting the law: platforms are not state regulators, and nor are they judges. On the other hand, however, since ToS are private agreements between user and platform that may unilaterally be changed at any moment in time, this creates a risk for legal certainty in terms of the forseeability of what is considered illegal online. This makes it difficult to discuss trends in illegal content online based on the Database, and therefore may not contribute to meaningful transparency.

Lastly, the Transparency Database is built on a standardization process undertaken by the European Commission. It is true that considerable efforts and resources have gone into setting up the enforcement framework of the DSA. The Commission is visibly increasing its legal and technological capacities in the process.[40] It is however debatable if these efforts are enough to bridge the considerable power and technology gap that currently exists at any level of governance between the public sector and VLOPs in particular. This knowledge asymmetry will inevitably affect what and how the Commission and national DSA enforcers will ask from online platforms. We already saw in our analysis that platforms are required to contextualize content moderation by providing original, free text insights into their practices, only for them to standardize basic 50 character sentences that do not add any value to a SoR. As we have explored in the compliance observations

[39] Laura Edelson, Inge Graef and Filippo Lancieri, 'Access to Data and Algorithms: For an Effective DMA and DSA Implementation' (Centre on Regulation in Europe 2023); Brandie Nonnecke and Camille Carlton, 'EU and US Legislation Seek to Open up Digital Platform Data' (2022) 375 Science 610.

[40] See for instance https://digital-strategy.ec.europa.eu/en/news/do-you-want-help-enforce-digital-services-act-apply-now-be-part-dsa-enforcement-team.

throughout Section 5, it is currently difficult to ascertain if many of these practices are actually compliant with the DSA. This uncertainty may weaken the Commission's negotiation stance with VLOPs in particular for further alignment with the objectives of the DSA.

Automated transparency mechanisms such as the DSA Transparency Database can certainly benefit from additional multidisciplinary attention. Future research should also consider comparisons between VLOPs and other online platforms which will also be under an obligation to submit SoRs as of 17 February 2024.

Appendices

a. SoR schema design

We provide the schema design of SoR in the Table 3.

Table: Schema of SoR designed by Commission as per Article 17 DSA. Fields marked * are mandatory.

Attribute	Description
uuid*	unique identifier for each SoR.
Content related attributes	
content_type*	type (text, image, video, etc) of content.
content_type_other	more details when content type is OTHER.
content_language	language of the content.
content_date*	date when content was posted on VLOP.
Type of restriction (visibility, monetary, provision of service, and account)	Any one of the fields marked with # has to be filled.
application_date	date when moderation decision was made by the VLOP.

Attribute	Description
decision_visibility#	restriction w.r.t. visibility can be categorized based on whether content is removed or disabled or age-restricted or others.
decision_visibility_other	to be filled when decision visibility doesn't fall into any pre-defined types like 'limited distribution', 'Video not eligible for recommendation in user's feed', etc.
end_date_visibility_decision	date when visibility restriction ends for the flagged content.
decision_monetary#	restriction (suspension or termination) w.r.t. monetary payments associated with the content being flagged.
decision_monetary_other	to be filled if other type of monetary restriction is being applied by VLOP which is not covered in pre-defined types.
end_date_monetary_restriction	date when monetary restriction ends for the flagged content.
decision_provision#	restriction w.r.t. provision of service (partial or total suspension, partial or total termination of service) to the recipient of service.
end_date_service_restriction	date when provision restriction ends for the flagged content.
decision_account#	restriction w.r.t. account being

Attribute	Description
	suspended or terminated.
end_date_account_restriction	date when account restriction ends for the flagged account.
account_type	type (business or private) of account.
territorial_scope*	it specifies the territorial scope of restriction.
Grounds on which decision is made, either incompatible or illegal content	
decision_ground*	reason for decision whether incompatible (if post is against terms of usage of a particular VLOP) or illegal content (if post is in violation of a national law).
decision_grounds_reference_url	link to the law or ToS of VLOP that is being violated.
incompatible_content_grounds (text)	reasons for content being incompatible which could include harassment & bullying, community guidelines, adult content, hate speech, violation of Terms of Service of VLOP, etc.
incompatible_content_explanation (text)	detailed explanations for reasons that content is incompatible.
illegal_content_legal_ground (text)	content is flagged illegal because it violates a national law.
illegal_content_explanation (text)	grounds of marking content illegal.

Attribute	Description
incompatible_content_illegal	it indicates whether moderated content was incompatible but also illegal.
category*	specifies the category of decision whether it is related to Terms of Service, illegal or harmful content, pornography, violence, unsafe products, and so on.
category_addition	any other additional category other than pre-defined list of categories.
category_specification	it specifies more keywords related to the category attribute.
category_specification_other	it is filled when KEYWORD_OTHER is specified in category specification.
decision_facts* (text)	it is a text field that gives details of the circumstances and facts relied upon in taking moderation decision.
source_type*	it specifies a pre-defined set of types of circumstances relied upon in taking moderation decision.
source_identity	it is optional field specifying notifier or source which helped in making moderation decision.
automated_detection*	it specifies whether automated means (yes or no) were used to make the detection.
automated_decision*	it specifies the extent to which (fully or partially automated or human

Attribute	Description
	intervention) automated means were used to derive at the decision.
created_at	date when SoR was submitted to the DTD database.

*Table 3. Schema of SoR designed by Commission as per Article 17 DSA. Fields marked * are mandatory.*

Chapter III: To Host or not to Host: Rethinking Intermediary Liability Exemption in the Digital Services Act

Published as: Jacob van de Kerkhof, 'To Host or not to Host: Rethinking Intermediary Liability Exemption in the Digital Services Act', 8 (2025) *Nordic Journal of European Law*, 79-120.

Abstract: *Intermediary liability exemption in the Digital Services Act (Regulation 2022/2065, 'DSA'), is generally interpreted as being unchanged compared the e-Commerce Directive (Directive 2000/31). Conditional liability exemption strikes a balance between the interests of hosting providers, people affected by illegal content, and the freedom of expression of the general public. However, hosting providers and their role in society have evolved significantly since the adoption of the e-Commerce Directive. Despite their position as a key infrastructure in the public debate and their sophisticated content moderation processes, they enjoy a seemingly widening exemption from liability in CJEU case law. This article argues for a narrower scope of application for intermediary liability exemption for hosting services guided by due diligence obligations and duties of care created in the DSA and sectoral regulation. This can benefit aggrieved parties in holding platforms liable and inspire better moderation of illegal content. A narrower interpretation of the liability exemption can be justified by compliance with requirements on notice-and-takedown mechanisms, a fairer economic burden of moderation, improved moderation capacity, and geopolitical tensions straining effective content moderation. The DSA provides clear due diligence obligations in article 16 and 23, and sectoral regulation requires diligence regarding specific types of content. These obligations can be used to exclude undiligent or bad faith actors from the liability exemption. This better reflects hosting providers' positions as powerful actors, and ensures a liability exemption that does not reward negligent or laissez-faire approaches to content moderation by actors that stand to gain monetarily from structurally hosting illegal content.*

Introduction

In February 2025, Instagram users were surprised by a large amount of violent and sexual imagery crowding their 'Reels' page. Meta apologised profusely for displaying illegal content to a large number of users: their automated moderation process had malfunctioned.[1] Platforms frequently boast the success of their content moderation systems, increasingly relying on novel techniques such as large-language models to moderate content;[2] the fallibility of such means, together with a changing political landscape in which platforms align with a political trend that promotes less moderation rather than more,[3] underline the importance of an effective liability framework surrounding the hosting of illegal content. A balanced intermediary liability framework strikes an appropriate balance between the interests of aggrieved parties, internet intermediaries, and the general public.

The primary liability exemptions for hosting illegal content in the European Union are laid down in articles 4-6 Digital Services Act ('DSA'[4]), and were previously found in articles 12-14 of the e-Commerce Directive ('ECD').[5] Intermediary liability for hosting illegal content constitutes a precarious balance between providing a remedy for parties affected by illegal content, protecting the users' freedom of expression from potential over-enforcement as a result of liability, chilling third-party speech, and stumping online platforms in their economic growth. The ECD's doctrine has been developed in case law over the last 25 years.[6] Academic consensus is that in the transition from the ECD to the DSA, intermediary liability exemptions have not

[1] 'Meta says it fixed 'error' after Instagram users report a flood of graphic and violent content', CBNC, https://www.cnbc.com/2025/02/27/meta-apologizes-after-instagram-users-see-graphic-and-violent-content.html

[2] For TikTok: Rozanna Latiff, 'ByteDance's TikTok Cuts Hundreds of Jobs in Shift towards AI Content Moderation' *Reuters* (11 October 2024) <https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai-moderation-sources-say-2024-10-11/> accessed 27 February 2025. For Meta: 'Content Moderation in a New Era for AI and Automation | Oversight Board' (17 September 2024) <https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/> accessed 27 February 2025.

[3] van de Kerkhof in Alberto Alemanno and Jacquelyn Veraldi, *Musk, Power, and the EU: Can EU Law Tackle the Challenges of Unchecked Plutocracy?* (Verfassungsbooks 2025) 91 et seq. <https://intrechtdok.de/receive/mir_mods_00019984> accessed 30 May 2025.

[4] 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).

[5] 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); see also Raphaël Gellert and Pieter Wolters, 'The Revision of the European Framework for the Liability and Responsibilities of Hosting Service Providers' (Ministry of Economic Affairs and Climate Policy 2021) 202006403 99.

[6] E.g. Court of Justice of the European Union, 23 March 2010, *Google France v Louis Vuitton*, C-236/08 & C-238/08; Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09; Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando*, C-682/18 and C-683/18; but also in the sphere of the Council of Europe: European Court of Human Rights, 16 June 2015, *Delfi v Estonia*, 64569/09; European Court of Human Rights, 15 May 2023, *Sanchez v France*, 45581/15; European Court of Human Rights, 7 January 2025, *Patrascu v Romania*, 1847/21.

changed.[7] The DSA has merely adopted some of the Court of Justice of the European Union's (CJEU) reasoning in its recitals to provide context to the liability exemption.[8] However, the landscape around content moderation and the dissemination of illegal content on the internet has developed significantly since 2000. Online platforms have become places where users socialise, network, participate in the democratic process, and shop, all at the same time.[9] The legal landscape has similarly evolved: the DSA creates a layer of accountability on top of the existing liability rules laid down in the ECD, encompassing a wide range of due diligence obligations, such as transparency obligations, risk assessments, and engagement with civil society actors.[10] These obligations follow a layered approach, increasing as the size of the platform increases.[11] Very Large Online Platforms and Very Large Online Search Engines ('VLOPSEs', platforms with a number of monthly active users exceeding 45 million)[12] need to abide by the most intense regime,[13] whereas micro- and small enterprises are excluded from the scope of a number of DSA provisions.[14]

The starting point of this contribution is the observation that the DSA maintains – at least in phrasing – the liability exemptions of the ECD in spite of the evolution that hosting providers underwent in the past 25 years. This raises the question whether changes in content moderation and the expectations raised in the DSA should be better reflected in how the doctrine of intermediary liability exemption is applied. This contribution argues that, although the legal phrasing of liability exemptions has not significantly changed since the ECD, it requires nuancing to (a) reflect platforms' poor compliance with mandatory notice-and-action mechanisms, frustrating a pre-

[7] E.g. Folkert Wilman, 'The Digital Services Act (DSA): An Overview' 2: 'Moreover, as will be seen below, the DSA includes the rules on liability which were previously contained in that directive.'; Aleksandra Kuczerawy, 'The Power of Positive Thinking Special Issues: Intermediary Liability as a Human Rights Issue' (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 226, 228 'After consultation, the Commission concluded that it would maintain the existing intermediary liability regime while implementing a sectorial, problem-driven approach.'; Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) 65 'To avoid [a death sentence for decentralised speech], the second generation of rules, such as the DSA, does not intend to change the underlying liability for someone else's content.'

[8] Ronan Riordan, 'A Case Study of Judicial-Legislative Interactions via the Lens of the DSA's Host Liability Rules' (2025) 2025 10 European Papers - A Journal on Law and Integration 259.

[9] Catalina Goanta, 'The New Social Media: Contracts, Consumers, and Chaos' (2023) 108 Iowa Law Review 118; generally Jose van Dijck, *The Culture of Connectivity: A Critical History of Social Media* (Oxford University Press 2013).

[10] Martin Husovec, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation Of Global Internet Rules' (2023) 38 Berkeley Technology Law Journal 883. See also Rachel Griffin, 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability' (2023) 14 Transnational Legal Theory 46; Robyn Caplan, 'Networked

[11] Regulation 2022/2065, rec. 76; see also F Wilman, 'De Digital Services Act (DSA): een belangrijke stap naar betere regulering van onlinedienstverlening' (2022) 28 Nederlands tijdschrift voor Europees Recht 220, 222.

[12] Regulation 2022/2065, art. 33(1).

[13] Regulation 2022/2065, Section 5, art. 33-43.

[14] As defined in Recommendation 2003/361/EC; Regulation 2022/2065, art. 15(2); art. 19(1), art. 29(1).

condition for liability exemption; (b) reflect the economic reality in which hosting providers benefit economically from hosting illegal content, (c) reflect platforms' improved technological moderation capacity compared to the early 2000s, (d) provide a clearer incentive for platforms to moderate illegal content in a geopolitical landscape that stimulates less moderation, and (e) the observation that the case law of the European Court of Human Rights regarding freedom of expression on the internet does not stand in the way of this argument. Not only can a narrower scope of liability exemption ensure a fairer distribution of responsibility for hosting illegal content, it provides more remedies for parties aggrieved by illegal content by excluding bad faith actors, and may stimulate moderation against illegal content,[15] preventing associated harms. The DSA provides clear guidelines that can guide this interpretation of the hosting liability exemption, predominantly through requirements that can exclude non-diligent hosting providers, and through expectations around the level of diligence that could lead to awareness of illegal content.

This argument is laid out along the following structure: Section I introduces the current framework for liability exemption for hosting illegal content and explains why the consensus is that the DSA has not changed the doctrine for liability exemption since the ECD; Section II provides reasons why the intermediary liability exemption should be interpreted more narrowly than 25 years ago; Section III explains instances in the DSA and sectoral regulation that can change how intermediary liability exemption is applied in light of requirements for platform due diligence. The purpose of this article is not to dismiss intermediary liability exemption; it argues for a nuanced narrowing of scope of the liability exemption in cases where hosting providers act negligently in their due diligence obligations in content moderation.

Section I: Intermediary Liability Exemption in the DSA

Before expanding on arguments for narrowing the scope of intermediary liability exemption in Section II, this section explains the framework for intermediary liability exemption for hosting illegal content in the DSA, and why it resembles that of the ECD. Important to note at the outset is that the DSA only harmonises the exemption of liability for hosting illegal content; generally, liability is established through national law or European law, and can be derived from criminal-, civil-, or any other

[15] E.g. Yassine Lefouili and Leonardo Madio, 'The Economics of Platform Liability' (2022) 53 European Journal of Law and Economics 319.

field of law.[16] The DSA harmonises a general horizontal liability exemption for mere conduit services, caching services, and hosting providers.[17] This article focuses exclusively on the exemption for hosting providers, because of their relevance in recent case law, and increased due diligence obligations in the DSA. This framework is explored in I.i. Section I.ii explains why the consensus is that the liability exemption has not changed since the ECD, and provides an overview of provisions that may stand in the way of interpreting the liability exemption differently. Section I.iii provides a brief interim conclusion.

I.i An overview of conditional intermediary liability exemption

The DSA's explanatory memorandum states that article 6 DSA maintains – at least textually – the liability exemption regime of article 14 ECD.[18] Article 14 ECD required Member States to ensure that hosting providers were not liable for hosting illegal content if they did not have knowledge of the illegal information they host.[19] The conditional liability exemption of Article 14 ECD was heavily inspired by para 512 of the U.S. Digital Millennium Copyright Act ('DMCA'),[20] a "sectorial, conditional immunity system."[21] The DMCA departed from the earlier US Section 230 Communications Decency Act,[22] which immunised hosting providers from liability for hosting most types of content.[23] Immunity in para 512 DMCA is conditional on the knowledge of the service provider of the copyright infringing content: if the service provider is unaware of copyright infringing content, they are not liable; once made aware of such content, the service provider has to act expeditiously to remove that content. Article 14 ECD copied this model for hosting providers.[24]

[16] Regulation 2022/2065, art. 3(h); Court of Justice of the European Union, 11 September 2014, *Papasavvas*, C-291/13, para 53; see also Folkert Wilman, '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' (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 317, 321.

[17] Regulation 2022/2065, art. 4-6.

[18] Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC; COM(2020) 825 final, p. 3. See also Franz Hofmann and Benjamin Raue (eds), *Digital Services Act: Article-by-Article Commentary* (First edition, Nomos 2025) 162.

[19] Directive 2000/31, art. 14(1)(a).

[20] 17 U.S.C. para 512.

[21] Husovec, 'Rising Above Liability' (n 10) 884.

[22] 47 U.S.C. para 230.

[23] See generally Jeff Kosseff, *The Twenty-Six Words That Created the Internet* (Cornell University Press 2019).

[24] Directive 2000/31, art. 14(1).

The popularity of liability exemption in the late 90s was based on the notion that less regulation would benefit economic interests and freedom of expression.[25] A framework for conditional liability exemption was seen as an agreeable middle ground between non-liability and strict liability. Non-liability would create an undesirable situation where parties affected by illegal content are left without remedies. Conversely, strict liability would create an undesirable scenario requiring hosting providers to assume full control over all content online, which would cause unpredictable monetary loss and diminish the benefits of a decentralised communication network.[26] If hosting providers were strictly liable for illegal content, they would need to exercise immense amounts of editorial control to avoid liability; this would limit the degree to which the internet can be used for accumulating free expression, especially since various courts recognise the internet's ability to provide a crucial vehicle for freedom of expression.[27] The downside of the liability exemption is that it leaves parties affected by illegal content with fewer remedies compared to a strict liability model, as hosting providers may hide behind being unaware of the illegal content. This is part of the balancing act that maintains the possibility of liability only if a hosting provider becomes aware of illegal content and precludes strict liability.[28] This is confirmed by the European Court of Human Rights ('ECtHR'), which held that the notice-and-action mechanism sufficiently safeguards affected parties' rights[29] – a finding that is further discussed below.[30]

The 'knowledge-based' liability exemption in article 6 DSA – and previously article 14 ECD – consists of three components:[31] (a) qualifying as a hosting provider; (b) not having actual knowledge of illegal content or as regards claims for damages, and not being aware of facts or circumstances from which the illegal content is apparent; or (c) acting expeditiously to remove or disable access to the illegal content.

---

[25] See also Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on A Digital Single Market Strategy for Europe, COM(2015) 192 final, p. 3-4.

[26] Husovec, 'Rising Above Liability' (n 10) 893.

[27] Court of Justice of the European Union, 26 April 2022, *Commission v Poland,* C-401/19, para 46; European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09, para 110; United States Supreme Court, 19 June 2017, *Packingham v North Carolina,* 137 S. Ct. 1730.

[28] E.g. Directive 2000/31, rec. 41 & 46; also Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 113.

[29] European Court of Human Rights, 2 February 2016, *MTE and Index.hu v Hungary,* 22947/13, para 91. *Cf.* European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09, para 159, in which the Court found that in some instances action is required even without notices.

[30] See *infra* Section III.i.a.

[31] Wilman, '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' (n 16).

I.i.a Hosting providers

Hosting providers are defined in art. 3(g)(iii) DSA: "a service consisting of the storage of information provided by, and at the request of, a recipient of the service."[32] They must take a neutral role, meaning that they provide merely technical and automated processing services, and not have control over user-generated content.[33] This criterium – in the context of article 14 ECD – has been interpreted extensively in case law of the CJEU that clarifies how a hosting provider may interact with information while maintaining a neutral role. For example, the CJEU established in *Google France* that remuneration for hosting services does not preclude a passive role.[34] Hosting providers may also engage in indexing, searching, and recommending user-generated content.[35] However, the CJEU decided in *L'Oreal v eBay*, that optimising the presentation of sales listings interfered with that neutral or passive role and excludes a hosting provider from its neutral position.[36] Neutrality of a hosting provider can therefore be seen as a spectrum encompassing a range of activities between full passivity and full control, that has developed over time along with platform affordances that must be assessed on a case-by-case basis.[37]

I.i.b Actual knowledge of illegal content or awareness of facts or circumstances from which illegal content is apparent

Hosting providers may not have 'actual knowledge' of illegal content, or be aware of any facts or circumstances from which illegal content is apparent.[38] These are seemingly two different standards, in which actual knowledge constitutes specific

[32] Regulation 2022/2065, art. 3(g)(iii); most obligations in the DSA are aimed at hosting services as opposed to mere conduit or caching services of arts. 3(g)(i) and (ii) respectively. Hosting providers have obligations under Chapter 3 Section 2; subspecies online platforms (art. 3(o)) have obligations under Section 3; online platforms allowing consumers to conclude distance contracts with traders under Section 4; and then VLOPs under Section 5; all are considered hosting providers.

[33] Regulation 2022/2065, art. 6(2), also rec. 18; Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 113. For a discussion whether neutrality also involves passiveness, see Folkert Wilman, 'Between Preservation and Clarification' (*Verfassungsblog*, 2 November 2022) <https://verfassungsblog.de/dsa-preservation-clarification/> accessed 8 May 2025.

[34] Court of Justice of the European Union, 23 March 2010, *Google France v Louis Vuitton,* para 116.

[35] Regulation 2022/2065, rec. 22; Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 114.

[36] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 116

[37] Joris van Hoboken and others, 'Hosting Intermediary Services and Illegal Content Online: An Analysis of the Scope of Article 14 ECD in Light of Developments in the Online Service Landscape.' (European Commission 2018) 31–33.

[38] It is important to note that rec. 18 seemingly conflates 'being a neutral hosting provider' with 'having knowledge or control over content', stating that "the provider of intermediary services plays an active role of such a kind as to give it knowledge of, or control over, that information." Although there is certainly overlap between 'having actual knowledge over' and 'being aware of', it is entirely conceivable that a passive or neutral hosting provider has knowledge over illegal content.

knowledge of the presence of illegal content and the nature of its illegality,[39] whereas awareness of facts or circumstances from which illegal content is apparent is a broader criterium that can be derived from a broader set of circumstances.[40] The difference is that in the case of 'awareness', liability can only lead to damages, whereas 'actual knowledge' could also lead to criminal liability.[41]

Actual knowledge or awareness of illegal content can be gained through own-initiative investigations or through notices submitted through notice-and-action mechanisms. In order to lead to awareness, notices must be sufficiently precise and adequately substantiated to allow a diligent economic operator to reasonably identify, assess, and where appropriate, act against illegal content.[42] The hosting provider must diligently assess the facts and circumstances brought forward in such notices.[43]

After a hosting provider has actual knowledge or awareness of allegedly illegal content, it must establish the illegal nature of the content.[44] The illegal nature must be "a matter of actual knowledge or be apparent", meaning that the illegality must be specifically established, readily identifiable, or manifest.[45] The findings of the CJEU on this are reflected in article 16(3) DSA. What is considered 'apparently' illegal can be contentious. A-G Saugmandsgaard Øe comments that in some cases illegality can be immediately obvious, mentioning child abuse material, whereas in others it may not be, mentioning copyright.[46] To avoid over-removal and hosting providers becoming judges of online legality, A-G Saugmandsgaard Øe argues that hosting providers can only be held liable for not removing content if the illegal character of the content can be ascertained without difficulty and without

[39] E.g. in the context of notices: Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Øe, C-682/18 and C-683/18, para 187: "A notification is also intended to give [the hosting provider] sufficient evidence to verify the illegal nature of the information."

[40] Husovec, *Principles of the Digital Services Act* (n 7) 130.

[41] ibid; Miquel Peguera, 'The DMCA Safe Harbors and Their European Counterparts: A Comparative Analysis of Some Common Problems' (2008) 32 Columbia Journal of Law and the Arts 481, 488–489.

[42] Regulation 2022/2065, rec. 22; art. 16(3).

[43] Regulation 2022/2065, art. 16(6)

[44] Court of Justice of the European Union, 23 March 2010, *Google France v Louis Vuitton,* C-236/08, C-237/08, C-238/08, para 109. See also Hofmann and Raue (n 18) 170.

[45] Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 113.

[46] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Øe, C-682/18 and C-683/18, para 187-190. Saugmandsgaard Øe argues that copyright cases can be difficult to assess due to the variety of rightholders that can be involved and the various national regimes that may need to be accounted for, referring explicitly to Court of Justice of the European Union, 16 February 2012, *SABAM v Netlog,* C-360/10, para 50. However, copyright cases have also been dubbed some of the more obvious cases of illegal content; in cases of piracy, illegality is often clear-cut.

conducting a detailed legal or factual examination i.e. when the illegal character is apparent.[47]

The standard for whether facts and circumstances have led to awareness of illegal content established in *L'Oreal/eBay* and in *YouTube/Cyando* is what a 'diligent economic operator' would have identified. This is a strict requirement that is not met easily: in *L'Oreal,* A-G Jääskinen excludes 'construed' knowledge on the basis of the diligence expected from an economic operator, meaning that a service provider cannot be liable on the basis of that they 'should have known', or whether they 'have good reasons to suspect illegal activity'.[48] Expecting otherwise could violate the prohibition on general monitoring obligations under article 15 ECD and 8 DSA. However, Jääskinen nuances this finding for 'same or similar' trademark infringing goods, arguing that 'same or similar' infringements are part of the same illegal activity, and as such constitute 'ongoing' illegal content, in which case the provider 'should have known'.[49]

The Court finds in *L'Oreal* that an imprecise or inadequately substantiated notice must be taken into account by national courts in establishing whether there are facts and circumstances from which a diligent economic operator should have identified the illegal content.[50] Nordemann interprets this as a duty of care, requiring some effort of the economic operator to identify illegal content based on facts presented to them.[51] *YouTube* nuances this finding. A-G Saugmandsgaard Øe seemingly dismisses the argument that awareness can be construed on the basis of the expected diligence or duty of care of an operator, stating: "attention should be paid not to the fact that the provider would have known had it been diligent, but to what it really knew."[52] The final decision in *YouTube* confirms this line of reasoning, finding that the notification must contain sufficient information to enable the operator to satisfy itself without a detailed legal examination that the content is illegal and that removing it is

[47] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Øe, C-682/18 and C-683/18, para 190.

[48] Court of Justice of the European Union, 9 December 2010, *L'Oreal v eBay,* Opinion of Advocate General Jääskinen, C-324/09, para 163.

[49] Court of Justice of the European Union, 9 December 2010, *L'Oreal v eBay,* Opinion of Advocate General Jääskinen, C-324/09, para 167.

[50] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 122.

[51] Jan Bernd Nordemann, 'Haftung von Providern im Urheberrecht Der aktuelle Stand nach dem EuGH-Urteil v. 12. 7. 2011 – C-324/09 – L'Oréal/eBay'.

[52] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Øe, C-682/18 and C-683/18, para 179; referring to Court of Justice of the European Union, 9 December 2010, *L'Oreal v eBay*, Opinion of Advocate General Jääskinen, C-324/09, para 163.

compatible with freedom of expression.[53] Thus, there is limited room in case law for a duty of care, and only in certain circumstances.

The level of diligence required from the 'diligent economic operator' is left relatively undefined unclear throughout this development. Adding to this complexity, Saugmandsgaard Øe links the liability exemption of article 14 ECD to the concept of good faith. Citing the input by the French Government, he finds that the liability exemption "seeks to protect service providers that generally act in good faith, not providers whose very intention is to facilitate copyright infringements."[54] Good faith can be demonstrated through generally complying with obligations to remove content, or by putting in place tools that combat illegal content.[55] However, judging by the tone of the CJEU[56] and the A-G,[57] it is interpreted primarily as a criterium of *not* deliberately facilitating illegal content, or remaining inactive despite having general awareness of the hosting service being used to disseminate illegal content.

I.i.c. Acts expeditiously to remove access to the illegal content

Upon having gained actual knowledge or awareness of illegal content, a hosting provider must act expeditiously to remove or disable access to the illegal content. This involves both a temporal requirement – content must be removed expeditiously – and the question of whether the platform could be required to remove content that is identical or similar to the illegal content.

The requirement of 'expedience' is undefined and needs to be assessed on a case-by-case basis.[58] It can be dependent on the nature of the content; for example, the expected expedience for removing CAM ('Child Abuse Material') may be greater than that of unwarranted financial advice. The requirement for member states is that temporal requirements may not be so rigid as to inspire over-blocking.[59] Previously, the *Conseil Constitutionnel* found that removal times under 24 hours in *Loi Avia* were

[53] Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 116, also Opinion of Advocate General Saugmandsgaard Oe, para 190.

[54] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, fn. 187.

[55] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, para 124 & 128.

[56] Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 94 on YouTube having various measures to prevent copyright infringement, and para 98, on Cyando not facilitating illegal sharing.

[57] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, para 191.

[58] Regulation 2022/2065, rec. 52; see also Hofmann and Raue (n 18) 178–179.

[59] Benjamin Raue and Hendrik Heesen, 'Raue/Heesen: Der Digital Services Act' [2022] Neue Juristische Wochenschrift 3537, 3540.

unconstitutional as they may lead platforms to remove content even in cases of doubt.[60] In contrast, the 2025 Code of Conduct on Countering Illegal Hate Speech Online+ requires signatories to review 50% of notices of illegal hate speech online within 24 hours.[61] The absolute time requirements posed by article 6(1)(b) DSA for takedown therefore remain somewhat vague, but in its Code of Conduct the Commission shows tolerance for stricter removal times.

Article 6(1)(b) DSA also raises the question whether awareness of specific illegal content could lead to an obligation to remove identical or similar content: can hosting providers be required to ensure *staydown* instead of *takedown*?[62] Staydown implies that future infringements must also be removed, which may require the hosting provider to moderate content *ex ante*. The notice-and-staydown model is popular in cases of copyright infringement: for example, under the German doctrine of *Störerhaftung* injunctions could be ordered for future infringing materials.[63] However, monitoring for similar infringements creates a tension with the general monitoring prohibition of article 8 DSA, as it would require hosting providers to proactively monitor content.

This tension is resolved somewhat inconsistently in CJEU case law over the past decade. In *Scarlet v SABAM* and *SABAM v Netlog*, the CJEU found that requiring preventive monitoring of all data by an internet service provider for intellectual property infringing content would contravene the general monitoring obligation of article 15(1) ECD.[64] In *McFadden* this reasoning was confirmed.[65] Conversely, in *L'Oréal*, the CJEU finds that injunctions to suspend violators of intellectual property rights may be ordered if operators of online marketplaces fail to take preventive measures, which encourages some forms of proactive content moderation.[66] The

[60] *Conseil Constitutionnel*, 18 June 2020, 2020-801 DC, para 16 see Ruth Janal, 'Haftung Und Verantwortung Im Entwurf Des Digital Services Acts' [2021] Zeitschrift für Europäisches Privatrecht 227, 248.

[61] Code of conduct on Countering Illegal Hate Speech Online + https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online, 2.3.

[62] Martin Husovec, 'The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown: Which Is Superior: And Why' (2018) 42 Columbia Journal of Law & the Arts 53. Also Aleksandra Kuczerawy, 'From "Notice and Takedown" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

[63] E.g. *Bundesgerichtshof,* 12 March 2004, *Internetversteigerung I,* I ZR 304/01; see for example Christina Angelopoulos and Stijn and Smet, 'Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability' (2016) 8 Journal of Media Law 266, 285; Kuczerawy, 'From "Notice and Takedown" to "Notice and Stay Down"' (n 62) 540.

[64] Court of Justice of the European Union, 24 November 2011, *Scarlet Extended v SABAM*, C-70/10, para 39-40. Court of Justice of the European Union, 16 February 2012, *SABAM v Netlog*, C-360/10.

[65] Court of Justice of the European Union, 15 September 2015, *Tobias McFadden v Sony Music Entertainment*, C-484/14, para 87.

[66] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 141.

CJEU nuances this line in *Glawischnig-Piesczek v Facebook*: it is not considered a general monitoring obligation when a hosting providers are required to remove identical – as already established in *L'Oreal* – and similar content.[67] The qualification 'similar' is a novelty. In *Glawischnig-Piesczek*, it is required by the nature of defamatory content, which may take different forms, but be similar at its core.[68] Similarity can only be established when it does not require an independent assessment by the hosting provider.[69] Janal argues that removing 'similar' content inevitably requires the use of filtering techniques, yet filtering may not lend themselves to effectively judging the context in which the defamatory speech is uttered. This may then contravene article 8 DSA.[70] To add a layer of proportionality to this requirement of filtering for similar content, A-G Saugmandsgaard Øe argues in his conclusion in *YouTube* that if monitoring obligations for identical or similar content are ordered, those obligations must be proportionate to the complexity of the obligation (i.e. similar content is more difficult to monitor than identical content), and therefore take into account the availability of monitoring systems to the hosting provider.[71] In sum, the CJEU does not fully dismiss the possibility of near-general monitoring obligations, but those need to be necessitated by the nature of the illegal content and proportional to the situation of the hosting provider.

I.i.d. Limitations to the scope of applicability of the liability exemption under article 6(1) DSA.

The DSA limits liability exemption compared to article 14 ECD in two instances: in article 6(3), and by leaving sectoral regulation on illegal content that creates different standards for liability exemption (e.g. the Copyright in the Digital Single Market Directive ('CDSM'[72])) intact.

Article 6(3) DSA stipulates that liability under consumer protection cannot be exempted for online platforms that allow the conclusion of distance contracts with traders where the average consumer could believe that the product was sold by the online platform itself.[73] This answers a critique that hosting providers are not liable for consumer harms even though consumers rely on the hosting provider's brand

[67] Court of Justice of the European Union, 3 October 2019, *Eva Glawischnig-Piesczek v Facebook,* C-18/18, para 37.

[68] *Id.,* para 40.

[69] *Id.,* para 46.

[70] Janal (n 60) 249–250.

[71] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, para 222.

[72] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (*OJ L 130,* 17 May 2019, p. 92).

[73] Regulation 2022/2065, art. 6(3).

image to buy goods and services.[74] Sectoral regulation fragments the landscape of intermediary liability exemption for illegal content outside of the DSA's scope. The Copyright Directive,[75] Enforcement Directive,[76] and CDSM Directive all act as *leges speciales* under the DSA.[77] Under the CDSM Directive online content-sharing service providers[78] perform an act of communicating to the public when hosting third-party content.[79] If they host copyright infringing content, they must obtain prior authorisation from right-holders to do so, or face liability for copyright infringement.[80] Providers may be exempt from liability if (a) they made best efforts to obtain authorisation, and (b) made, in accordance with high industry standards of professional diligence, best efforts to ensure the unavailability of specific works, and (c) acted expeditiously to remove content upon receiving a notification and made best efforts to remove future uploads.[81] This is a higher standard of liability exemption than article 6 DSA, which is justified by addressing right-holders from a value-gap created by negligent enforcement on behalf of online service providers.[82] To be exempt from liability in the case of copyright, online platforms may be required to employ automated filtering techniques,[83] creating a tension with article 8 DSA and generally the right to freedom of expression.[84] The CJEU has resolved this tension in *Poland v Commission*, finding that platforms are not obliged to generally

[74] E.g. Caroline Cauffman and Catalina Goanta, 'A New Order: The Digital Services Act and Consumer Protection' (2021) 12 European Journal of Risk Regulation 758, 766–767. Caufmann and Goanta criticise the provision (then 5(3) of the DSA proposal) for creating legal uncertainty on the interpretation of average consumer – although that is a common standard in consumer law, it is still unclear what factors lead an average consumer to assume that the hosting provider is party to the contract

[75] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (*OJ L 167*, 22 June 2001, p. 10).

[76] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights (*OJ L 157*, 30 April 2004, p. 45).

[77] Regulation 2022/2065, rec. 11.

[78] Directive 2019/790, art. 2(6); this can be seen as a subset of online platforms in the sense of art. 3(i) DSA.

[79] Directive 2001/29, art. 3.

[80] Directive 2019/790, art. 17(1); Directive 2001/29/EC, art. 3.

[81] Directive 2019/790, art. 17(4).

[82] Directive 2019/790, rec. 66. *Cf.* on whether this justifies stricter liability than in other fields, Janal (n 60) 235–236.

[83] Christophe Geiger and Bernd Justin Jütte, 'Platform Liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match' (2021) 70 GRUR International 517, 550–552.

[84] This provision has been popular topic of academic debate, see for example: Geiger and Jütte (n 83); Joao Pedro Quintais and others, 'Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive Statements' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 277; Martin Husovec, 'Mandatory Filtering Does Not Always Violate Freedom of Expression: Important Lessons from Poland v. Council and European Parliament' (2023) 60 Common Market Law Review 173; Martin Senftleben and Christina Angelopoulos, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the e-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' [2020] Amsterdam/Cambridge.

monitor, and that article 17(7)-(9) provides sufficient safeguards to balance copyright and the right to freedom of expression.[85]

I.ii Why is the consensus that the hosting liability exemption has not changed?

As iterated above, the academic consensus is that the framework for liability exemption has not changed since the ECD. The fundamental reasoning behind the liability exemption – balancing the promotion of economic development in internet services, maintaining freedom of expression by not mandating general monitoring for all illegal content, and offering users a remedy via notice-and-action and potential liability – remains intact.[86] Most reports evaluating the intermediary liability exemption generally find that, although the ECD required clarification, there was no real support for fundamentally altering the liability exemption.[87] The liability exemption regime was considered 'fit-for-purpose' by most individual users, content uploaders, and intermediaries.[88] Only IP-rightholders indicate flaws in both the system and its interpretation in case law, stating it does not allow them sufficient redress from intermediaries and infringing parties. Their concerns have been addressed in Directive 2019/790.[89]

Instead, the DSA adds a layer of accountability norms in the form of due diligence and transparency, which are not linked to liability. Husovec calls this 'accountability-but-not-liability.'[90] These provisions require additional efforts from platforms, but do not translate to intermediary liability. This separation may stand in the way of a more narrow interpretation of the liability exemption in the DSA compared to the ECD. This subsection describes some provisions that would stand in the way of a more narrow interpretation of the liability exemption.

---

[85] Court of Justice of the European Union, 26 April 2022, *Commission v Poland*, C-401/19, para 72-97.

[86] Generally, Wilman, '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' (n 16) 322–323.

[87] E.g. van Hoboken and others (n 37); Gellert and Wolters (n 5); Tambiama Madiega, 'Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In Depth Analysis.' (European Parliament 2019) PE 649.404; Anja Hoffmann and Alessandro Gasparotti, 'Liability for Illegal Content Online' (Centre for European Policy 2020).

[88] European Commission, 'Synopsis report on the public consultation on the regulatory environment for platforms, online intermediaries and the collaborative economy', 2015, accessible via: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15877; see also Maria Lillà Montagnani, 'A New Liability Regime for Illegal Content in the Digital Single Market Strategy' in Giancarlo Frosio (ed), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

[89] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (*OJ L 130,* 17 May 2019, p. 92). See section 1.iii; also more extensively João Pedro Quintais and Sebastian Felix Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?' (2022) 13 European Journal of Risk Regulation 191.

[90] Husovec, 'Rising Above Liability' (n 10) 910. See also Giancarlo Frosio and Martin Husovec, 'Accountability and Responsibility of Online Intermediaries' in Giancarlo Frosio (ed), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

The first provision anchoring liability exemption amidst other norms is article 54. Article 54 explicitly separates liability for hosting illegal content from liability for individual harms suffered by infringement of the DSA.[91] The reasoning behind this distinction is that liability for illegal content is aimed at restoring a lawful state by making victims whole, whereas liability under article 54 is only about holding platforms accountable for complying with their obligations under the DSA.[92] Accountability provisions in the DSA on transparency and due diligence are therefore not intended create liability for the communication of others.[93] Recital 41 underscores this sentiment: "the due diligence obligations are independent from the question of liability of providers of intermediary services, which therefore need to be assessed separately."[94] To clarify this separation with an example: it is impossible to hold a platform for hosting CAM because it was late in submitting transparency reporting.[95] Since a separate avenue for liability for the novel obligations stemming from the DSA's due diligence system is available, it is not necessary to link the doctrine for intermediary liability exemption to the due diligence system.

A second provision preventing a more limitative reading of liability exemption is the Good Samaritan provision of article 7.[96] Article 7 confirms that intermediary service providers are not excluded from relying on liability exemptions of articles 4-6 DSA if they carry out voluntary good faith investigations into illegal content or take necessary measures to comply with national- or EU law, if this leads to an active or non-neutral role that would normally exclude them from qualifying as a hosting provider. This provision prevents a catch-22 in which an online platform would be excluded from the liability exemption because they moderate content in good faith.[97] Exclusion from liability exemption for good faith content moderation could lead to less moderation to avoid liability.[98] Good faith implies that platforms operate in an objective, non-discriminatory and proportionate manner and take into account the

[91] Regulation 2022/2065, art. 54.

[92] Husovec, 'Rising Above Liability' (n 10) 911.

[93] Husovec, *Principles of the Digital Services Act* (n 7) 66.

[94] Regulation 2022/2065, rec. 41.

[95] Hofmann and Raue (n 18) 958–960.

[96] Regulation 2022/2065, art. 7.

[97] van Hoboken and others (n 37) 39. This fear was not justified even without the Good Samaritan provision, see e.g. Directive 2000/31, rec. 40; European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Tackling Illegal Content Online, COM(2017)555 final, 28 September 2017, p. 12-13; see also Gellert and Wolters (n 5) 68–69.

[98] By analogy: "where ignorance is bliss, 'tis wise to be folly", California Court of Appeals, June 23 1958, *People v Smith*, Civ. A. No. 3792.

rights and legitimate interests of all the parties involved.[99] The application of the Good Samaritan may be limited by the requirement of diligence, critics argue. Berbec, referring to Kuczerawy, questions whether the liability exemption is available, for example, if a hosting provider only succeeds in removing some illegal content on the basis of a notice, but misses some identical or similar content of which the illegal nature is apparent.[100] An undiligent hosting provider may therefore be excluded from the provision, leading to the conclusion that "in other words, taking voluntary actions in good faith neither guarantees no[r] precludes neutrality."[101] In spite of this, the article is commonly seen as an assurance that platforms should moderate content without fearing lability, and therefore ensures that the liability exemption also extends to hosting providers that actively moderate content.[102]

A final limitation is laid down in article 8 DSA, which prohibits general monitoring obligations.[103] The ECD allowed for member states to create explicit duties of care regarding certain types of illegal content.[104] The DSA does not copy this, making the general monitoring prohibition stricter.[105] General monitoring obligations are prohibited due to operational costs for hosting providers and risks to the freedom of expression. The use of automated filtering is associated with risks to freedom of expression because of the potential for catching proverbial 'dolphins in the net', where legal content is inadvertently removed.[106] Article 8 must be interpreted as

---

[99] Regulation 2022/2065, rec. 26, see also Jacob van de Kerkhof, 'Good Faith in Article 6 Digital Services Act (Good Samaritan Exemption) – The Digital Constitutionalist' (15 February 2023) <https://digi-con.org/good-faith-in-article-6-digital-services-act-good-samaritan-exemption/> accessed 7 April 2025.

[100] Adriana Berbec, 'To What Extent Can Online Service Providers Adopt Voluntary Content Moderation Measures without Losing Their Immunity Shields? A Comparative Analysis of Online Service Providers' Liabilities in the European Union and the United States' (2024) 15 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law; Aleksandra Kuczerawy, 'The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act' (*Verfassungsblog*, 12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 21 February 2025.; the requirement to moderate against content of which the illegal nature is apparent arises from Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando*, Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, para 197.

[101] Kuczerawy, 'The Good Samaritan That Wasn't' (n 100).

[102] Wilman, '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' (n 16) 335. Unclear: Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 109.

[103] Regulation 2022/2065, art. 8; previously included as art. 15 Directive 2000/31.

[104] Directive 2000/31, rec. 48.

[105] Arguably, the explicit duties of care are included in sectoral regulation, see e.g. Gerald Spindler, 'The Liability System of Art. 17 DSMD and National Implementation: Contravening Prohibition of General Monitoring Duties' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 344; Senftleben and Angelopoulos (n 84).

[106] E.g. Daphne Keller, 'Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion' (2019) <https://cyberlaw-stanford-edu.proxy.library.uu.nl/files/Dolphins-in-the-Net-AG-Analysis.pdf> accessed 25 November 2021; more generally: Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.

precluding any level of diligence being expected of an operator that may lead to a *de facto* or *de jure* general monitoring obligation or pro-active measures relating to identifying or monitoring illegal content.[107] Specific monitoring obligations are allowed for similar or identical content, however, following *Glawischnig-Piesczek*.[108] *De facto* general monitoring obligations are created by assuming a duty of care that is too expansive. Article 8 DSA can therefore stand in the way of limiting the liability exemption: by making hosting providers more liable, providers are *de facto* expected to assume a more active role in content moderation, which is precluded by article 8.

I.iii: Intermediary findings

This section has presented an overview of the intermediary liability exemption in article 6 DSA. Neutral intermediary service providers are not liable for hosting illegal content if they have no specific knowledge of that content or its illegality, and if they act expeditiously to remove it after gaining knowledge. The liability exemption for hosting providers is justified by facilitating economic growth of digital service providers and ensuring that freedom of expression is not interfered with through over-removing content by hosting providers fearing liability. Despite the added responsibility for hosting providers codified in the DSA and sector-specific regulation, the liability exemption has not significantly changed since the e-Commerce Directive of 2000. Its scope is somewhat limited by article 6(3) DSA and sectoral regulation, but its key characteristics remain. In fact, altering the scope of the liability exemption in the DSA is prevented by the Good Samaritan provision of article 7 DSA, the separate ground for liability of article 54 DSA, and the general monitoring prohibition of article 8 DSA. However, this article argues that there is room for a nuanced rethinking of the intermediary liability exemption guided by provisions in the DSA. Section II outlines arguments that justify narrowing of the scope of intermediary liability exemption within the framework of the DSA. This rethinking does not dismiss the key characteristics of the liability exemption, but ensures that it applies to good faith actors whilst excluding bad faith hosting providers.

Section II: Why should we rethink the intermediary liability exemption of article 6 DSA?

This article contests that DSA allows for liability exemption to be interpreted more narrowly than it was under the ECD, at least on a case-by-case basis. In some

[107] Regulation 2022/2065, rec. 22.

[108] Court of Justice of the European Union, 3 October 2019, *Eva Glawischnig-Piesczek v Facebook,* C-18/18.

instances, this could aid individual aggrieved parties and rightholders with an additional remedy vis-à-vis the platform, which is currently precluded by a broad liability exemption. Aside from the general observation that laws applying to the internet drafted in the year 2000 may not be entirely fit for a fast-developing digital economy, this contestation rests on five points: (i) insufficient access to functioning notice-and-action mechanisms warrants a narrower application of the liability exemption; (ii) a narrower exemption would reflect a fairer division of the economic burden of moderation; (iii) a narrower liability exemption reflects developments in moderation capacity over the past two decades; (iv) a narrower exemption could contravene a geopolitical interest in less moderation and create enforcement mechanisms against mis-moderation; (v) case law of the European Court of Human Rights indicates that the freedom of expression does not necessarily preclude a stricter liability exemption.

II.i Insufficient access to functioning notice-and-action mechanisms

Notice-and-action mechanisms of online platform providers constitutes a cornerstone in the intermediary liability exemption structure outlined in I.i. Liability exemption can – at least under the ECD and DSA – only be justified if parties aggrieved by illegal content have a remedy available. That remedy is provided by an avenue for the affected party to request takedown of the illegal content. This is usually achieved through a notice-and-action mechanism. The importance of such mechanisms is acknowledged by the EU regulator;[109] the ECtHR in *MTE v Hungary*, finding that a functioning notice-and-action mechanism acts as a balancing instrument between various interests, and would thus not justify liability for hosting comments;[110] and the CJEU in acknowledging the importance of notice-and-action mechanisms in combatting copyright infringements in *Youtube/Cyando*.[111] National case law also emphasises the importance of such mechanisms in facilitating users to act against illegal content.[112]

Previously, the requirements for notice-and-action were fragmented on a member state level.[113] The DSA progresses this thorny issue in harmonising those

[109] European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Tackling Illegal Content Online, COM(2017)555 final, 28 September 2017, p. 16.

[110] European Court of Human Rights, 2 February 2016, *MTE and Index.hu v Hungary,* 22947/13, para 91.

[111] Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 94.

[112] E.g. Hoge Raad, 27 January 2023, *Brein v NSE,* ECLI:NL:HR:2023:94, para 1.4 et seq.

[113] Aleksandra Kuczerawy, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (2015) 31 Computer Law & Security Review 46, 51.

requirements, explicitly acknowledging that (a) hosting providers must provide their users with an easily accessible and user-friendly means of reporting content;[114] and (b) that a notice through such as mechanism creates knowledge or awareness of illegal content with the hosting provider, thus excluding them from liability exemption if they do not act expeditiously.[115]

The design of such mechanisms greatly affects how users are able to submit notices. Several studies have indicated that hosting providers do not adhere to requirements of ease-of-access or user-friendliness. In works on Germany's *Netzwerkdurchsetzungsgesetz*, which also included a notice-and-takedown requirement, Heldt[116] and Wagner et al.[117] show that Facebook's design of their notice-and-action mechanism affects how users report content. The reporting affordance for illegal content was considerably less accessible than that for incompliance with terms and conditions, resulting in users submitting less notices on illegal content. Sekwenz et al[118] and Holznagel[119] observe a similar trend regarding article 16 DSA notice-and-action mechanisms for TikTok and Facebook: notice-and-action mechanisms are poorly implemented, leading to users being dissuaded from reporting illegal content, or not being able to report it as illegal. Article 16 has led to various complaints against VLOPs.[120] If the avenue for generating knowledge or awareness of illegal content with the hosting provider is dysfunctional, hosting providers remain within the scope of the liability exemption, since it is more difficult to create awareness of illegal content. Non-compliance with article 16 can therefore, to some extent, be rewarding; the user is less able to alert the platform and therefore trigger liability. The only remedy currently for individuals against this is article 54 DSA. However, if functioning notice-and-action mechanisms

---

[114] Regulation 2022/2065, art. 16(1).

[115] Regulation 2022/2065, art. 16(3).

[116] Amélie Heldt, 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports' (2019) 8 Internet Policy Review.

[117] Ben Wagner and others, 'Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <https://dl.acm.org/doi/10.1145/3351095.3372856> accessed 18 October 2023.

[118] Marie-Therese Sekwenz, Ben Wagner and Simon Parkin, '"It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!" – Evaluating Reporting Mechanisms under the Digital Services Act', *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2025).

[119] Daniel Holznagel, 'Follow Me to Unregulated Waters!' [2024] Verfassungsblog <https://verfassungsblog.de/follow-me-to-unregulated-waters/> accessed 23 September 2024.

[120] E.g. TikTok: hateaid, 'HateAid Files a Complaint against TikTok' (*HateAid*, 14 February 2025) <https://hateaid.org/en/systemic-failure-hateaid-files-complaint-against-tiktok/> accessed 26 May 2025. Also Meta: 'Commission Opens Formal Proceedings under DSA' (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373> accessed 26 May 2025.

are a precondition for liability exemption, it seems plausible that not providing such mechanisms could lead to exclusion from the liability exemption, as is argued in Section III.i.a.

II.ii Fairer economic burden for hosting illegal content

Narrowing liability exemption can be justified because it requires internet intermediaries, who generally benefit economically from hosting illegal content, to be accountable to parties seeking a remedy when affected by illegal content. Buiten et al have conducted a law and economics analysis of liability rules for hosting services.[121] Their analysis starts with the observation that hosting providers are stuck in a paradox where they may want to act against illegal content online, but simultaneously profit from the presence of that content on their services. Since they benefit economically from hosting content, it is fair to place responsibility for the costs of content moderation with those hosting providers. Their economic benefit can also be used as an argument for holding them liable: "[i]f hosting platforms reap the benefits from exchange in the space they govern, it stands to reason that they also face responsibility if their business model causes harm".[122] Their analysis is supported by the so-called 'value-gap'. The 'value gap' is a term coined by copyright holders to describe the situation in which the hosting provider generates income over infringing content but the original rightsholder does not.[123] The CDSM Directive addresses the value gap for copyright holders through a stricter liability regime.[124] However, the internet economy has developed beyond generating income through copyright infringing content, but other content remains under a 'loose' liability exemption regime. Users – and therefore hosting providers – also generate income by defaming people, or spreading hatred.[125] The negative effects – material and immaterial – of such speech are not carried by the platform, however; they are carried by citizens. It could be considered a fairer distribution of the burden for hosting illegal content if bad faith actors profiting monetarily from hosting illegal

[121] Miriam C Buiten, Alexandre De Streel and And Martin Peitz, 'Rethinking Liability Rules for Online Hosting Platforms' (2020) 28 International Journal of Law and Information Technology 139.

[122] ibid 149.

[123] Annemarie Bridy, 'The Price of Closing the "Value Gap": How the Music Industry Hacked EU Copyright Reform' (2020) 22 Vanderbilt Journal of Entertainment & Technology Law 323.

[124] Martin Senftleben, João Pedro Quintais and Arlette Meiring, 'How the EU Outsoucres the Task of Human Rights Protection to Platforms and Users: The Case of UGC Monetization' (2023) 38 Berkely Technology Law Journal 101, 105.

[125] Thales Bertaglia, Catalina Goanta and Adriana Iamnitchi, 'The Monetisation of Toxicity: Analysing YouTube Content Creators and Controversy-Driven Engagement', *4th International Workshop on OPEN CHALLENGES IN ONLINE SOCIAL NETWORKS* (ACM 2024).

content would be excluded from liability exemption, allowing aggrieved parties to hold them liable.

II.iii Improved moderation capacity

Content moderation processes have improved leaps since the early 2000s due to advances in automation. Automation is necessitated by the enormous scale on which content is moderated.[126] Hosting providers have since developed sophisticated tools to automate content moderation.[127] To illustrate: every 6 months, roughly 10 billion content moderation decisions are uploaded in to the DSA Transparency Database, the vast majority of them automated.[128] Academic discourse has focused primarily on weaknesses in automated procedures regarding freedom of expression[129] and transparency;[130] however, it is undeniable that content moderation has advanced significantly, especially since human moderation is also not an absolute guarantee of accuracy. It is expected that advancements in the use of large language models in content moderation will add to content moderation accuracy in the future.[131]

The increased accuracy of content moderation since the early 2000s leads to higher detection and higher removal rates of illegal content; a development that is boasted by the largest social media platforms, who increasingly rely on automated moderation at the expense of human moderation.[132] Their boasting of successful moderation processes also raises expectations. The expectation of accuracy is underlined in several (co-)regulatory documents, for example the Code of

[126] Tarleton Gillespie, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 205395172094323.

[127] For example: 'Working Document Impact Assessment on the Modernisation of the EU Copyright Rules' (European Commission 2016) Commission Staff Working Document SWD(2016) 301 final.

[128] Amaury Trujillo, Tiziano Fagni and Stefano Cresci, 'The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media' (2025) 9 Proc. ACM Hum.-Comput. Interact. CSCW187:1; Rishabh Kaushal and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024).

[129] E.g. Federica Casarosa, 'When the Algorithm Is Not Fully Reliable: The Collaboration between Technology and Humans in the Fight against Hate Speech' in Hans-W Micklitz and others (eds), *Constitutional Challenges in the Algorithmic Society* (1st edn, Cambridge University Press 2021); Celine Castets-Renardt, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) 2020 University of Illinois Journal of Law Technology & Policy 283.

[130] E.g. Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973; Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18.

[131] E.g. Emmanuel Vargas Penagos, 'ChatGPT, Can You Solve the Content Moderation Dilemma?' (2024) 32 International Journal of Law and Information Technology eaae028.

[132] Latiff (n 2); 'Meta's New AI System to Help Tackle Harmful Content' (*Meta Newsroom*, 8 December 2021) <https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/> accessed 21 May 2025.

Conduct+[133] against illegal hate speech and the Regulation on terrorist content online.[134] Although article 7 DSA precludes such good faith investigations leading to liability (see section I.ii), the expected accuracy created by social media networks' own PR, as well as the evolvement of content moderation into a sophisticated process, should factor in assessing liability exemption under article 6. Not only will the advancement of content moderation lead to awareness of illegal content on a factual basis more frequently, it also raises the question whether platforms choosing not to apply such measures should be held liable for negligence.[135] Of course, it is a conscious choice of the EU regulator not to mimic the liability exemption of the CDSM Directive. However, the DSA's liability exemption stems from a period in time when it was impossible for hosting providers to control all content. Even though this is even more true nowadays due to the exponential growth of hosting services, automation and sophistication do allow hosting providers to better control content. This requires rethinking whether hosting providers should be exempt from liability even if they have advanced methods available – especially in cases where they choose not to use such methods.

II.iv Providing a countermovement to a trend of moderation withdrawal

As illustrated by the anecdote in the introduction, it is not uncommon for intermediary service providers to make mistakes in content moderation. This is not problematic *per se*, most platform regulations acknowledge that perfect moderation is impossible.[136] However, the current political climate in the United States supports reducing content moderation efforts. Protected by the geopolitical influence of the Trump administration, which is publicly threatening the EU if it applies the DSA too stringently,[137] CEO's of online platforms increasingly move towards less moderation.[138] This means that the dissemination of illegal content online increases;

[133] Code of conduct on Countering Illegal Hate Speech Online + https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online, art. 4.2.4.

[134] Regulation 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, *OJ L 172*, p. 79-101 (TERREG), rec. 16.

[135] See by analogy Directive 2019/790, art. 17(4).

[136] See e.g. Castets-Renardt (n 129).

[137] The White House, 'Fact Sheet: President Donald J. Trump Issues Directive to Prevent the Unfair Exploitation of American Innovation' (*The White House*, 21 February 2025) <https://www.whitehouse.gov/fact-sheets/2025/02/fact-sheet-president-donald-j-trump-issues-directive-to-prevent-the-unfair-exploitation-of-american-innovation/> accessed 31 May 2025.

[138] Dean Jackson Szóka Berin, 'The Far Right's War on Content Moderation Comes to Europe | TechPolicy.Press' (*Tech Policy Press*, 11 February 2025) <https://techpolicy.press/the-far-rights-war-on-content-moderation-comes-to-europe> accessed 7 April 2025.

as an example, the prevalence of hate speech on X has increased significantly since Musk became its CEO.[139]

Intermediary liability can be an effective incentive for hosting providers to remove illegal content more actively. Liability needs to be weighed against the risks to freedom of expression posed by over-moderation. The Commission and the CJEU, through case law and subsequent codification in the DSA have arguably opted for a broad scope for intermediary liability exemption. Nevertheless, narrowing the interpretation of the intermediary liability exemption in some cases can be justified to inspire better moderation efforts by intermediary service providers to provide counterweight to the 'moderation withdrawal'.

II.v Freedom of expression is no *prima facie* obstacle to a more narrow interpretation

A main point of critique against restricting intermediary liability exemption is that it would interfere with freedom of expression, since strict liability may lead to over-removal. However, the ECtHR leaves considerable room for intermediary liability in its case law, suggesting that freedom of expression does not stand in the way of a more narrow interpretation of the liability exemption.

In its case law, the ECtHR balances rights of individuals affected by defamation or hate speech against the right to freedom of expression of hosting providers. In that balancing act it does not consider the liability exemption of article 14 ECD or 6 DSA[140] – nor does it have to.[141] For example, in *Delfi v Estonia*, the ECtHR found that holding a newspaper liable for hate speech posted by anonymous users was not in violation of the right to freedom of expression of article 10 ECHR, without considering that such a newspaper would enjoy liability exemption.[142] In *MTE v Hungary*, the ECtHR considered holding a Hungarian association liable for hosting

[139] Daniel Hickey and others, 'X under Musk's Leadership: Substantial Hate and No Reduction in Inauthentic Activity' (2025) 20 PLOS ONE e0313293.

[140] European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09, para 127; Oreste Pollicino, 'European Judicial Dialogue and the Protection of Fundamental Rights in the New Digital Environment: An Attempt at Emancipation and Reconciliation' in Sonia Morano-Foadi and Lucy Vickers (eds), *Fundamental Rights in the EU: A Matter for Two Courts* (Bloomsbury Publishing 2015) 108; Marta Maroni, 'The Liability of Internet Intermediaries and the European Court of Human Rights' in Bilyana Petkova and Tuomas Ojanen, *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020) 268.

[141] Martin Husovec and others, 'Grand Confusion after Sanchez v. France: Seven Reasons for Concern about Strasbourg Jurisprudence on Intermediaries' (2024) 31 Maastricht Journal of European and Comparative Law 385. *Cf.* Michael FitzGerald, 'Not Hollowed by a Delphic Frenzy: European Intermediary Liability from the Perspective of a Bad Man: A Response to Martin Husovec et Al.' (2025) 32 Maastricht Journal of European and Comparative Law 47.

[142] European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09, para 160-161. Lisl Brunner, 'The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after Delfi v Estonia' (2016) 16 Human Rights Law Review 163.

'false and offensive' comments to interfere with the freedom of expression, since the notion of liability applied by the national courts did not allow for a balancing of interests between competing rights.[143] It explicitly did not consider the liability exemption.[144] In its latest seminal judgement on the matter, *Sanchez v France,* the ECtHR confirms the line of *Delfi*, and even extends it to apply to a politician who was held criminally liable for not removing hateful comments by third parties on his Facebook page.[145] In this case too, the Court did not rely on the liability exemption following the national court's assessment on that matter,[146] but also emphasises that "to exempt producers (as a specific subclass under French civil law) from all liability might facilitate or encourage abuse and misuse, including hate speech and calls to violence, but also manipulation, lies and disinformation."[147]

The argument here is that freedom of expression does not need to stand in the way of a more narrow interpretation of the liability exemption. When interpreting the DSA in light of the Charter[148] – which forms a guiding document following recital 153,[149] the case law of the ECtHR leaves room for a more narrow interpretation of liability exemption without interfering with freedom of expression. This narrow interpretation would emphasise the need for due diligence by hosting providers. Of course, the potential risks to freedom of expression of stricter liability need to be weighed against the benefits of that liability, and much depends on the factual assessment by national courts of whether the internet intermediary qualifies as a hosting provider. The ECtHR has developed a line of case law that explicitly acknowledges the possibility for establishing liability for internet intermediaries, and the freedom of expression does not necessarily need to stand in the way of stricter liability exemption.

This section has provided arguments for a more narrow interpretation of intermediary liability exemption. These are rooted in the conviction that, since the role of hosting providers in society has increased significantly since the year 2000, and their capacities have evolved correspondingly, it is fair to come to a more narrow interpretation of the liability exemption. Of course, this narrow

[143] European Court of Human Rights, 2 February 2016, *MTE and Index.hu v Hungary,* 22947/13, para 89.

[144] European Court of Human Rights, 2 February 2016, *MTE and Index.hu v Hungary*, 22947/13, para 51; it accepted the national court's reasoning that MTE did not qualify as a hosting provider because it did not provide e-commerce services.

[145] European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 209.

[146] European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 137-139.

[147] European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 185.

[148] Which must be interpreted in light of the European Court of Human Rights case law following art. 52(3) Charter of Fundamental Rights of the European Union.

[149] Regulation 2022/2065, rec. 153.

interpretation needs to be grounded in law. The next section explores avenues in the DSA that allow for a more narrow interpretation of the intermediary liability exemption.

Section III: How does the DSA allow for rethinking the intermediary liability exemption?

Section II has provided arguments for why a narrow interpretation of the hosting liability exemption may be desirable and possible from a fundamental rights perspective. This section explores which provisions in the DSA can be used to that argumentation. The DSA can allow applying the liability exemption more narrowly. The purpose of the narrow interpretation is to exclude bad faith actors from the liability exemption. An actor can be considered in bad faith when it is generally incompliant with the DSA's due diligence norms, or undiligent in its handling of notices to avoid awareness of illegal content. Although section I.ii has explained why due diligence obligations are separated from liability questions, the threshold for intermediary liability exemption still introduces a level of diligence expected from intermediaries to determine whether they are a hosting provider that was not aware of illegal content.

This raises a question: what happens when an undiligent economic operator is presented with facts or circumstances that would have created awareness of illegal content within the framework of the DSA had they been diligent? This section argues that an undiligent economic operator may be excluded from the liability exemption for two reasons: (i) intermediary services that structurally infringe on article 16(1) and 23(1) DSA do not qualify as a neutral hosting provider in the sense of article 6(1) DSA, and/or (ii) the undiligent hosting provider has failed to remove content of which the diligent provider ought to have been aware in light of article 16(6) DSA and sectoral due diligence norms.

III.i. An undiligent intermediary can be excluded from qualifying as a neutral hosting provider under article 6(1)

A first line of argumentation is that an undiligent intermediary service provider should not qualify as a passive, neutral hosting provider. The requirements for a hosting provider are that internet intermediary services provide their services neutrally, by mere technical and automatic processing of information.[150] To be excluded from this category, a hosting provider either has active knowledge over

[150] Regulation 2022/2065, rec. 18, see Section 1.1.a.

content, or has control over the content shared. This threshold is not met easily: as mentioned in I.i.a, platforms may structure, recommend, and index content without being excluded from being a hosting provider. However, as demonstrated in I.i.b, the hosting provider may not deliberately facilitate dissemination of illegal content: hosting providers are excluded from liability exemption when they have openly expressed the intention to facilitate illegal content.[151] This is also reflected in recital 20 of the DSA: in instances where the intermediary service "deliberately collaborates" with a user they are not neutral and therefore should not benefit from the liability exemption. In line with the case law cited above, this is the case when intermediary service providers make explicit that their purpose is to facilitate illegal activities.

However, the threshold of open expression of intention to facilitate illegal content is not insignificant, and is therefore unlikely to exclude many 'bad faith' hosting providers, even though that could be desirable from the perspective of combatting illegal content and providing remedies to aggrieved parties. Holznagel notes that the requirements of neutrality and passivity, the 'knowledge-or-control' test, does not necessarily exclude providers that are structurally prone to hosting illegal content (*'strukturell gefahrgeneigter Provider'*) from liability exemption.[152] He mentions Darknet platforms and platforms with limited protection for revenge porn as examples.[153] These types of hosting providers may be passive, neutral and not control the illegal content hosted on their platform, and thus qualify as a hosting provider under article 6 DSA, exempting them from liability if they remove content expeditiously after a notice.[154] At the same time, they are prone to hosting illegal content. This protection may therefore be overly broad: although platforms cannot be held liable based on general knowledge that they host illegal content under CJEU case law, it seems overly protective to offer liability exemption to platforms who have such general knowledge and act non-diligently in the design of their services with respect to illegal content. An exemption from liability may not encourage any

[151] See to that effect: Court of Justice of the European Union, 26 april 2017, *Stichting Brein v Filmspeler*, C-527/15; Court of Justice of the European Union, 14 June 2017, *Stichting Brein v Ziggo & XS4All*, C-610/15; Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando*, Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, para 120;129; & 191.

[152] Again, with the exception of those that openly facilitate illegal content, in this case *Filmspeler* and *ThePirateBay*; Daniel Holznagel, 'Nach dem EuGH-Urteil in Sachen YouTube/Cyando: Fast alles geklärt zur Host-Provider-Haftung?' (2021) 37 Computer und Recht 603, 605.

[153] Daniel Holznagel, 'Chapter II Des Vorschlags Der EU-Kommission Für Einen Digital Services Act – Versteckte Weichenstellungen Und Ausstehende Reparaturen Bei Den Regelungen Zu Privilegierung, Haftung & Herkunftslandprinzip Für Provider Und Online-Plattformen' (2021) 37 Computer und Recht 123, 124.

[154] Regulation 2022/2065, art. 6(1)(b).

moderation efforts, as Buiten et al have stipulated, and may even be conducive to sinister business-models that seek to monetise illegal content.[155]

A-G Saugmandsgaard Øe concluded -albeit in a footnote- in 2020 that the liability exemption is only intended for hosting providers acting in good faith.[156] This consideration has not been adopted explicitly in decisions by the CJEU, or in the DSA.[157] Holznagel wrote in 2021 that excluding bad faith hosting providers as 'non-neutral' would require a formal alteration of the standards for being a hosting provider.[158] However, the distinction between passive and active has diminished somewhat in the DSA; the DSA explicitly allows platforms to take an active stance toward content without losing their liability exemption, for example by indexing or filtering content.[159] An answer to the impossibility regarding the exclusion of bad faith actors would be to exclude structurally prone-to-illegal-content platforms from qualifying as a neutral[160] hosting provider, and therefore excluding them from the liability exemption.

What does it mean to be a non-neutral hosting provider? Under the ECD, being neutral -aside from the practical implications of providing the service without taking any editorial responsibility or control over the content- was generally interpreted as abiding by takedown requests and taking some content moderation measures, but left room for interpretation on where the borders of neutrality are.[161] It can be argued that under the DSA, some aspects are formally required from neutral hosting providers in order to maintain their liability exemption: article 9 DSA provides a clear avenue establishing clear takedown requests by administrative entities, which, by repeated refusal or non-compliance, could lead to the conclusion that the hosting provider is not neutral. Similarly, article 16(6) requires hosting providers to engage in good faith with notices submitted through notice-and-action mechanisms. Equally, article 23, on abusive users, requires that hosting providers act in good faith in maintaining their user-base, and exclude abusive recipients from their service. If a

[155] Buiten, De Streel and Peitz (n 121) 150–151.

[156] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* Opinion of Advocate General Saugmandsgaard Oe, C-682/18 and C-683/18, fn. 187.

[157] Although article 7 DSA on the Good Samaritan provision could be read as rewarding good faith economic operators with liability exemption, even if they are aware of illegal content.

[158] Holznagel, 'Nach dem EuGH-Urteil in Sachen YouTube/Cyando' (n 152) 606.

[159] Regulation 2022/2065, rec. 22.

[160] Focussing on neutrality rather than passivity is in line with CJEU case law, e.g. Court of Justice of the European Union, 9 December 2010, *L'Oreal v eBay,* Opinion of Advocate General Jääskinen, C-324/09, para 138-146; see also van Hoboken and others (n 37) 31–32.

[161] Wilman, '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' (n 16) 321.

significant part of the user base of a platform is abusive, this could exclude the hosting provider from neutrality. Non-compliance could preclude hosting provides from acting in a 'neutral' capacity, and therefore exclude them from the liability exemption of article 6. The subsections III.i.a and III.i.b explain why articles 16 and 23 provide legal instruments that enable the exclusion of intermediary service providers that are structurally prone to hosting illegal content or generally act in bad faith from the liability exemption.

III.i.a Dysfunctional or no notice-and-action mechanisms under Article 16

As noted in II.i, notice-and-action mechanisms are an important aspect of the intermediary liability exemption framework. However, before the DSA, there was no general legal requirement for intermediary services to provide such a mechanism.[162] Varying national implementations, and some sectoral (co)-regulations recommended,[163] or indeed required,[164] intermediary services providers to provide such mechanisms. Article 16 adds to this paradigm with a horizontal requirement for hosting providers to provide user-friendly and easily-accessible mechanisms that allows anyone to notify them of illegal content.[165] Remedies for users for non-compliance with this article would, in line with the framework outlined in I.iv, fall under article 54 DSA on damages.

However, conceptualising the hosting provider as a neutral, good faith intermediary, it can be argued that failing to provide users with a notice-and-action mechanism under article 16(1) DSA excludes intermediary service providers from qualifying as a hosting provider, and thus from liability exemption. Holznagel underscores this argumentation, noting that missing notice-and-action mechanisms could be an indicator for an 'active role' due to creating a structural risk for illegal content.[166] The reasoning for this is that notice-and-action is the only available remedy for users in

[162] Pieter Wolters and RaphaeI Gellert, 'Towards a Better Notice and Action Mechanism in the DSA' (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 403; MZ Van Drunen, 'The Post-Editorial Control Era: How EU Media Law Matches Platforms' Organisational Control with Cooperative Responsibility' (2020) 12 Journal of Media Law 166, 177. Although some argue that the ECD indirectly requires platforms to adopt notice and action mechanisms, see Aleksandra Kuczerawy, 'From "Notice and Take Down" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' in Giancarlo Frosio (ed), *The Oxford Handbook of Intermediary Liability Online* (Oxford University Press 2019).

[163] E.g. Code of Conduct on Countering Illegal Hate Speech Online, accessible via https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en; now updated to the Code of Conduct+ on Countering Illegal Hate Speech Online, adopted under art. 45 DSA.

[164] Regulation 2021/784, art. 10; Directive 2019/790, art. 17(9).

[165] Regulation 2022/2065, art. 16(1); article 3(h).

[166] Holznagel, 'Chapter II Des Vorschlags Der EU-Kommission Für Einen Digital Services Act – Versteckte Weichenstellungen Und Ausstehende Reparaturen Bei Den Regelungen Zu Privilegierung, Haftung & Herkunftslandprinzip Für Provider Und Online-Plattformen' (n 153) 126.

the liability exemption framework, with the exception of overly burdensome and time-consuming procedures to acquire injunctions in national courts, which still may be an ineffective remedy due to the speed at which illegal content spreads on the internet.[167] If intermediary service providers do not present users with the possibility of 'creating' awareness of illegal content – after all, a notice under article 16 is seen as creating such awareness[168] – they might not be able to be held liable. Lacking a notice-and-action mechanism should be seen as bad-faith design of the hosting service; by being negligent in respect of article 16, the hosting provider ensures that liability exemption of article 6(1)(a) would always be available. This cannot be the intention of the EU regulator. Not facilitating a notice-and-action mechanism must therefore lead to a potential exclusion from the intermediary liability exemption.

Following this argumentation, the question arises whether the degree of non-compliance with article 16 should factor into the assessment whether the hosting provider takes an active approach. Notice-and-action mechanisms, after all, can take different forms,[169] and compliance with the requirements of article 16 is varying across hosting providers.[170] This has to be assessed on a case-by-case basis, in line with the spectrum of neutral activities mentioned in I.i.a. It is disproportional to the ratio of article 6 to exclude intermediary service providers from liability exemption for being slightly difficult to use; however, if the notice-and-action mechanism is absent, or is designed in such a way that users are unable to meet the criteria of submitting an adequately substantiated notice by using these mechanisms, it could be fair to exclude them from being labelled as a neutral hosting provider. For example, if a platform does not facilitate an explanation of the illegality of a notice – it is possible that this precludes an adequately substantiated notice, which requires an explanation of the legal ground under which the content is illegal and an explanation of why the content is illegal.[171] Being unable to explain the illegality in sufficient detail would leave the user unable to create awareness under CJEU case law.

---

[167] That is not to say that users cannot attempt to request takedown outside of the mechanisms of article 16; see recently Kammergericht Berlin, 25 August 2025, 10 W 70/25.

[168] Regulation 2022/2065, art. 16(3).

[169] Kate Crawford and Tarleton Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 New Media & Society 410.

[170] Sekwenz, Wagner and Parkin (n 118).

[171] See further examples: Marie-Therese Sekwenz, Ben Wagner and Simon Parkin, '"It Is Unfair, And It Would Be Unwise To Expect The User To Know The Law!" – Evaluating Reporting Mechanisms Under The Digital Services Act' <https://data.4tu.nl/datasets/16efb52d-b94d-4180-a5f6-89d93028a8c2> accessed 21 August 2025.

A functioning-notice-and-action system is a *conditio sine qua non* for the liability exemption framework in the DSA. Rewarding intermediary service providers that fail to contribute to that system with liability exemption is unfair, especially in light of the many (co-)regulatory instruments requiring such a mechanism. The proportionality of excluding hosting providers from the liability exemption is ensured by the fact that article 16(1) poses form-free requirements, and hosting providers have the freedom to comply in a way that suits their service best.[172] The requirements therefore do not significantly impede on the freedom of enterprise of hosting providers, nor do they pose excessive burdens.[173] It is therefore not unfair that incompliance with article 16 should lead to the hosting provider being excluded from the liability exemption of article 6 DSA, by excluding such providers from qualifying as a neutral hosting provider, especially in light of the strict requirements posed on notices to create specific knowledge or awareness of illegal content.[174]

III.i.b Misuse of services under article 23(1)

Article 23(1) is a due diligence obligation that requires online platforms to suspend users that frequently provide manifestly illegal content.[175] It requires online platforms (as a subcategory of hosting providers[176]) to be aware of users that misuse their services and exclude them. This prevents online platforms from being populated with illegal content by 'repeat offenders'. The argument in this subsection is that structurally disregarding abusive users that frequently post illegal content online may exclude an intermediary service provider from qualifying as a neutral hosting provider. The CJEU acknowledged that repeat infringers could require an additional level of diligence by the hosting provider. In *L'Oreal,* the Court decided that hosting providers may be required to suspend users who repeatedly post copyright infringing content to prevent future infringements.[177] Hofmann underscores this argument, noting that infringing on the duty of care laid down in article 23 can lead to the assumption of an active role for the platform provider.[178]

---

[172] Hofmann and Raue (n 18) 328.

[173] cf. Wolters and Gellert (n 162) 408; Christophe Geiger, Giancarlo Frosio and Elena Izyumenko, 'Intermediary Liability and Fundamental Rights' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020).

[174] Regulation 2022/2065, rec. 22: "[…] through notices submitted to it by individuals or entities in accordance with this Regulation in so far as such notices are sufficiently precise and adequately substantiated to allow a diligent economic operator to reasonably identify, assess, and where appropriate, act against the allegedly illegal content. Also: Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* C-682/18 and C-683/18, para 116.

[175] Regulation 2022/2065, art. 23(1).

[176] Regulation 2022/2065, art. 3(i).

[177] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 141.

[178] Hofmann and Raue (n 18) 163.

The *ratio* behind the exclusion from the liability exemption is that structural non-compliance with article 23 is an indication of a bad faith approach to content moderation. The indication that non-compliance represents bad faith is informed the fact that article 23 in itself represents high thresholds regarding the frequency and the manifest illegality of the content disseminated, meaning that only a negligent hosting provider could have missed an abusive user. The frequency necessary to trigger article 23(1) is dependent on the requirements laid down in art. 23(3): the absolute volume of illegal content provided, and the proportion of the illegal content compared to legal content.[179] The manifest illegality of the content is dependent on whether a layperson would see the content as evidently illegal.[180] It could be argued that the nature of the manifestly illegal content should be a factor in determining compliance with article 23: while a layperson may be able to detect fake Birkenstocks as trademark infringement, a platforms failure to block accounts spreading CAM may be more damning under article 23(1) if the requirement refers to 'manifestly illegal content'. Similarly, one could argue that frequency and manifest illegality can be corresponding factors, meaning that a high frequency of less evidently illegal content could still trigger article 23(1), as well as a comparatively lower frequency of very evidently illegal content, such as CAM. Since excluding online providers from the liability exemption is a significant step, the proportionality of the exclusion needs to be weighed against the degree of non-compliance with article; it would be disproportional to exclude a platform with over a billion users due to failing to suspend services for one abusive user. Non-compliance with article 23 therefore needs to be structural. This needs to be assessed on a case-by-case basis and can depend *inter alia* on the nature of the illegality prevalent on the platform, the proportion of users that can be labelled as abusive under article 23, and the measures taken to combat such abuse.

Aside from not qualifying as a neutral hosting provider, it could be argued that, even if the online platform was unaware of illegal content posted by such a 'repeat offender', it should have been due to the requirements of article 23(1). This relates to the point made in the next subsection in which provisions in the DSA can be used to lower the threshold for when a hosting provider is deemed aware of illegal content. In the case of abusive users, knowledge of that illegal content is then construed by

[179] Regulation 2022/2065, art. 23(3).

[180] Regulation 2022/2065, rec. 63. This seemingly constitutes a stricter barrier than the illegality that needs to be indicated by users in a notice under article 16, as found in: Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando,* C-682/18 and C-683/18, para 116: "sufficient information to enable the operator of that platform to satisfy itself, without a detailed legal examination, that that communication is illegal and that removing that content is compatible with freedom of expression."

the explicit expectations of the due diligence provision in article 23(1). Again, this provision needs to be applied proportionally to the impact that excluding hosting providers from liability exemption would have. The bar for being an abusive user in article 23(1) is therefore not met easily, Raue argues, and only with respect to the interests of all parties involved.[181] This may limit the instances where non-compliance with article 23(1) leads to exclusion from intermediary liability exemption in practice. However, the high threshold of applying article 23(1) underscores the severity of the non-compliance by the hosting provider, and supports the notion that it is proportional that when a platform provider is taking a bad-faith approach to content moderation, and fails to take action in article 23(1) situations, they no longer qualify as a neutral hosting provider.

III.ii The threshold for awareness in diligent economic operators is lower

The previous subsection has argued that non-compliance with article 16 and 23 DSA may lead to exclusion from qualifying as a neutral hosting provider; those actors can therefore not rely on the liability exemption by default. This subsection moves one step further: if an intermediary service provider qualifies as a hosting provider, provisions in the DSA and sectoral regulation can lower the threshold of awareness of illegal content with diligent economic operator under article 6(1)(a)DSA in factual assessments. If the hosting provider does not act expeditiously after becoming aware of illegal content, they are excluded from the liability exemption following article 6(1)(b).[182]

The expectation of diligence in relation to awareness of illegal content is introduced in Section I.i.b: a hosting provider may not rely on the liability exemption if they are aware of facts or circumstances that would have led a diligent economic operator to identify the illegality. This criterion, to some extent, creates an expectation of diligence from hosting providers. This diligence can be interpreted as a duty of care. The CJEU created some room for this interpretation in *L'Oréal*, which was not completely resolved in *YouTube*, especially in light of the tension that a duty of care would bring with the prohibition on general monitoring obligations. In *L'Oréal*, the CJEU found that an imprecise notice must still be taken into account by a diligent operator.[183] *YouTube* clarifies that in those instances the notice must still contain sufficient information to remove the content without a detailed legal examination.[184]

[181] Hofmann and Raue (n 18) 472.

[182] Regulation 2022/2065, art. 6(1)(b).

[183] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 122.

[184] Court of Justice of the European Union, 16 July 2020, *YouTube/Cyando*, C-682/18 and C-683/18, para 116.

The degree to which this requires a duty of care was never fully resolved in case law, but *Youtube* indicates that the duty of care is not broad. However, the DSA introduces a number of provisions that can be used to define the threshold for when a hosting provider is deemed aware of illegal content in line with that duty of care.

This section proposes that developments in the DSA and sectoral regulation arguably lower the factual threshold for establishing awareness of illegal content on a factual level compared to the ECD, for example in cases where a notice is insufficiently precise, or even absent. This is supported by two arguments: (a) article 16(6) explicitly requires better diligence on notices on ongoing infringement; (b) duties of care established in sectoral regulation reduce the amount of information required to raise knowledge or awareness. These two arguments are supported by the general notion that, since content moderation mechanisms have become incrementally more sophisticated than they were under the ECD (see Section II.iii), it is not unreasonable to expect that hosting providers have awareness of illegal content sooner than they would have 25 years ago.

III.ii.a: Article 16(6) DSA requires more diligence in assessing notices

As part of the newly formalised notice-and-action framework, article 16(6) DSA stipulates that decisions on the information in any notices received through the mechanisms of article 16(1) must be taken in a timely, diligent, non-arbitrary and objective manner.[185] This is predominantly a codification of jurisprudence, and outlines requirements that can be used to fill in a 'duty of care' for hosting providers on notices. It is important to note that these principles are a matter of procedure, not outcome. Article 16(6) does not entitle a user to a correct decision, just to a timely, non-arbitrary, objective, and diligent decision.[186] Such decisions may also be rendered using automatic means.[187] Aside from the timeliness requirement, which is fluid and needs to be established on a case-by-case basis, notices need to be examined in a diligent, non-arbitrary and objective manner. While non-arbitrariness and objectivity are requirements that are part of the proceduralisation of content moderation,[188] the argument of this subsection is that the requirement of diligence of

[185] Regulation 2022/2065, art. 16(6).

[186] Wolters and Gellert (n 162) 415.

[187] Explicitly: art. 16(6): "[…] *Where they use automated means for that processing or decision-making*, they shall include information on such use in the notification referred to in paragraph 5." [emphasis added]

[188] Interestingly they assume that the hosting provider is a neutral actor that has no opinions or preferences of its own. However, the European Court of Human Rights acknowledged to some extent in *Google v Russia* that hosting providers can have their own freedom of expression: see European Court of Human Rights, 8 July 2025, *Google LLC and others v Russia,* 37027/22, para 91.

article 16(6) may lower the threshold for awareness of illegal content, as a diligent economic operator can be deemed aware of illegal content sooner.

The next question is: how diligent does a notice need to be examined? Generally, the level of diligence required is relatively low, Holznagel argues, in light of the finding in *Youtube* that illegality needs to be established without a detailed examination of the notice.[189] However, providers could be assumed to have gained knowledge if the provider "negligently ignores evidence for a violation."[190] On a practical level this includes information included in a notice by users, but also existing and easily accessible knowledge, such as the context of the alleged violation, the history of the reporter, and the history of the uploader.

In line with the argumentation in II.iii: it is not inconceivable that existing and easily accessible knowledge is more present than it was 20 years ago; it could therefore be argued that factually, negligently ignoring evidence of a violation is likely to happen more frequently. The DSA also creates a number of examples that facilitate access to knowledge outside of the information contained in the notice. For example, trusted flagger status, as part of the history of the reporter, under article 22 can be a relevant factor.[191] On the history of the uploader: if a user frequently uploads manifestly illegal content, the online platform is expected to be aware of that content under article 23. As argued in the previous subsection, knowledge on illegal content can be construed in such situations based on the explicit due diligence obligation for the online platform. Previous knowledge on the nature of the infringing content could be established through earlier notices under article 16, appeals under article 20,[192] or potentially earlier decisions regarding identical or similar content in the out-of-court dispute settlement framework of article 21.[193] It is possible that the same content is reported multiple times by the same reporter: the DSA does not prevent *ne bis in idem* with regards to notices on illegal content. This causes a hosting provider to have some prior knowledge on a specific infringement.

An additional question is whether article 16(6) requires diligence only with regards to the specific information contained in the notice, or also requires reflection on

[189] Daniel Holznagel, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (2024) 25 Computer Law Review International 172, 177.
[190] ibid.
[191] Regulation 2022/2065, art. 22(1). Arguably, the opposite is true for users that frequently abuse the notice-and-action mechanism. The online platform is expected to suspend their flagging mechanisms under art. 23(2).
[192] Regulation 2022/2065, art. 20.
[193] Regulation 2022/2065, art. 21.

similar or identical illegal content, thus creating a potential takedown obligation for more than the content exactly indicated in the notice. This tension was touched upon under section I.ii.c. as notice-and-staydown.[194] Holznagel argues that this is not the intention of the notice-and-action framework of article 16(6): the relating recitals and legislative process do not support it.[195] However, if identical or similar content is understood as an ongoing infringement, as opposed to separate information that was not included in the original notice, then a diligent hosting provider might be expected to address such content.

The understanding of ongoing infringement is in line with A-G Jääskinen's conclusion in *L'Oreal*, in which he illustrates that: "if A has been discovered infringing trade mark X by listing an offer on the electronic marketplace in September, I would not exclude that the marketplace operator could be considered having actual knowledge of information, activity, facts or circumstance if A uploads a new offer of the same or similar goods under trade mark X in October. In such circumstances it would be more natural to speak about the same continuous infringement than two separate infringements. I recall that Article 14(1)(a) mentions 'activity' as one object of actual knowledge. An ongoing activity covers past, present and future."[196] Jääskinen's argumentation includes a strong temporal dimension (September … October). Whether this applies to a wider timeframe (e.g. September … January) needs to be assessed on a case-by-case basis. It is difficult to speak of ongoing infringement if this timeline was 5 years instead of a mere month; such would be overly burdensome to the hosting provider, and create a tension with the prohibition on general monitoring obligations. However, qualifying similar- and identical content as ongoing infringement may require a hosting provider to remove such illegal content as well as part of the diligence required under article 16(6). This expected diligence can create circumstances in which a platform is deemed aware of illegal content, even though a notice referring the content may not have been specific enough to raise knowledge on illegal content. This interpretation of article 16(6) would be in line with the CJEU's decision in *Glawischnig-Piesczek*, in which an effective remedy could only be achieved through granting an injunction for 'similar' content. If a hosting provider fails to address content that *should* have been identified through diligent assessment of the notice, it could be excluded from article 6(1). Of

[194] Court of Justice of the European Union, 3 October 2019, *Eva Glawischnig-Piesczek v Facebook,* C-18/18, para 41.

[195] Holznagel, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (n 189) 178.

[196] Court of Justice of the European Union, 9 December 2010, *L'Oreal v eBay,* Opinion of Advocate General Jääskinen, C-324/09, para 167.

course the level of diligence in this case needs to be balanced against the general monitoring prohibition, but *Glawischnig-Piesczek* opens the door to a more expansive reading of diligence.

III.ii.b: Duties of care established in sectoral regulation and the DSA raise the bar for the diligent economic operator

The DSA is the horizontal cornerstone of the digital *acquis*, but succeeds a number of sectoral (co-)regulations that establish duties of care for hosting providers (or a sector-specific subset thereof). Some sectoral regulation exists outside of the scope of DSA liability exemption, such as the CDSM.[197] However, other regulatory and co-regulatory instruments, such as the amendments to the Audiovisual Media Services Directive (AVMSD),[198] the Regulation preventing the dissemination of terrorist content online (TERREG),[199] and several codes of conduct,[200] introduce explicit due diligence obligations for hosting providers or their corresponding sector-specific subset. This subsection argues that the level of diligence of economic operators required by article 6(1) DSA needs to be interpreted in line with duties of care that are imposed on hosting providers in the varying sectoral regulations. In cases where information is not clearly available through the notice submitted because it is incomplete, duties of care required in sectoral (co-)regulation may lower the threshold of information contained in the notice for the platform to still make an assessment on the legality of the content.

One example of sectoral regulation requiring additional due diligence from hosting providers is the amended AVMSD. The DSA applies without prejudice to this Directive.[201] The AVMSD applies to media service providers – a categorisation distinct from the categorisations made in the DSA. Particularly relevant are 'video-sharing platform services', a subset of media service providers, are essentially platforms that allow the dissemination of user-generated videos to the general

[197] Regulation 2022/2065, rec. 11; Directive 2019/790.

[198] Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, *OJ L 303*, 28.11.2018, p. 69–92 (AMSD).

[199] Regulation 2021/784.

[200] Code of Conduct on Disinformation, https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation; Code of conduct on Countering Illegal Hate Speech Online + https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

[201] Regulation 2022/2065, rec. 10. Interestingly article 28b applies without prejudice to articles 12 to 15 of the ECD, which leaves the relation between the DSA and the AMVSD unclear.

public.[202] This affordance is common for online platforms under the DSA; most VLOPs and online platforms are therefore required to comply with the AVMSD parallel to the DSA.[203] Under article 28b(1) AVMSD, video-sharing platforms are expected to adopt appropriate measures to protect minors from commercial content that impairs their development, and the general public from commercial content containing violence, hate speech, terrorist content, or CAM.[204] The appropriateness of such measures is determined in light of the nature of the content, the harm it may cause, and the interests of the group protected and the public at large.[205] This generally means that larger platforms will need to take more appropriate measures against such content than smaller platforms.[206] Platforms must include and apply such measures in their terms and conditions – underlining that some level of diligence is required.[207] This is safeguarded by establishing a notice-and-action mechanism and dispute resolution procedures.[208] Appropriate measures are bound by the prohibition of general monitoring obligations of article 8, and can therefore not involve preventive filtering of content.[209] Nevertheless, Ullrich underlines that some "morally and technically founded duties of care" would be appropriate in the context of the AVMSD.[210] This gap is, since his writing, partially filled by the due diligence framework of the DSA, but the notion that the requirements of the AVMSD create an extra duty of care for platforms still stands after the DSA's adoption. A more diligent investigation into notices regarding CAM, hate speech, or terrorist content is therefore required from video-sharing platform services by reading the AVMSD in conjunction with the DSA.

Another example of sectoral regulation introducing additional requirements for hosting providers is TERREG. TERREG applies parallel to the DSA and the AVMSD.[211] It applies to hosting service providers, a definition found in the

[202] Directive 2010/13, art. 1(aa).

[203] Sally Broughton Micova and Ľuboš Kukliš, 'Responsibilities of Video-Sharing Platforms and Their Users' in Heritiana Ranaivoson, Sally Broughton Micova and Tim Raats (eds), *European Audiovisual Policy in Transition* (Routledge 2023) 88.

[204] Directive 2010/13, art. 28(b).

[205] Directive 2010/13, art. 28b(3)

[206] Pietro Dunn, 'Online Hate Speech and Intermediary Liability in the Age of Algorithmic Content Moderation' (University of Bologna 2024) 90–91.

[207] Directive 2010/13, art. 28b(3)(a)-(b), also Giovanni De Gregorio, 'Expressions on Platforms': (2018) 2 European Competition and Regulatory Law Review 203, 212–213.

[208] Directive 2010/13, art. 28b(3)(d)-(i).

[209] Directive 2010/13, art. 28b(3); *mutatis mutandis*: "Those 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."

[210] Carsten Ullrich, 'Standards for Duty of Care: Debating Intermediary Liability from a Sectoral Perspective' (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 111, 118–119.

[211] Regulation 2021/784, art. 1(5); Regulation 2022/2065, rec. 10.

Information Society Services Directive that largely mirrors the definition of hosting provider in the DSA.[212] The Regulation tackles the dissemination of terrorist content online by laying down duties of care for hosting providers.[213] When a hosting provider is confronted with terrorist content in the last 12 months – which is virtually all large online platforms – they are required to take extra measures.[214] These measures include addressing removal orders within one(!) hour, flagging channels for EUROPOL and other competent authorities, notice and action mechanisms, and taking proactive measures to combat illegal content.[215] Several researchers have criticised this Regulation for introducing a pro-active duty of care that would contravene article 8 DSA.[216] Platforms would be expected to use automated means to filter illegal content, which arguably presents risks for freedom of expression, certainly in light of the broad definition of terrorist content in the Regulation.[217] This critique is generally well-founded; however, with regards to liability exemption – which van Hoboken argues is partly undermined by the Regulation – it could be argued that platforms are expected to be more diligent when users submit notices on terrorist content, both regarding the required expedience in taking action and the degree to which platforms are expected to take down identical and similar content. Because the platform is required in sectoral regulation to be more diligent, the threshold for when a platform is aware of that illegal content is lower when a user reports terrorist content, even when that notice is not precise, or if it did not remove content identical to a previous notice.

Sectoral codes of conduct can also add requirements to content moderation practices, in particular with regards to illegal hate speech and disinformation.[218] The DSA explicitly formalises Codes of Conduct as part of the DSA framework.[219] Codes of Conduct are traditionally voluntary, multistakeholder efforts that shape policy and facilitate participation in content moderation.[220] Although the DSA emphasises their

[212] Regulation 2021/784, article 2(1); Directive 2015/1535, art. 1(b).

[213] Regulation 2021/784, art. 1(1)(a).

[214] Regulation 2021/784, art. 4(4).

[215] Regulation 2021/784, art. 4(1).

[216] *Inter alia* Alexandre De Streel, 'Online Platforms' Moderation of Illegal Content Online' (European Parliament 2020) PE 652.718 25–26; Joris van Hoboken, 'The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications' (Transatlantic Working Group on Content Moderation Online and Freedom of Expression 2019) 7.

[217] Regulation 2021/784, art. 2(7).

[218] https://digital-strategy.ec.europa.eu/en/policies/dsa-codes-conduct

[219] Regulation 2022/2065, art. 45-47; this contribution addresses only article 45.

[220] Rachel Griffin, 'Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns' (2024) 2024 Technology and Regulation 167, 179.

voluntary nature,[221] it has been theorised that there is significant pressure to adhere to the standards of such codes.[222] This is underscored by recital 104 stressing that "[…] the refusal without proper explanations by a provider of an online platform or of an online search engine of the Commission's invitation to participate in the application of such a code of conduct could be taken into account, where relevant, when determining whether the online platform or the online search engine has infringed the obligations laid down by this Regulation. […]"[223] Generally, complying with codes of conduct will make hosting providers more aware of illegal content on a practical level, for example by engaging with trusted flaggers and fact-checkers.[224] These interactions should generally not lead to exclusion of the liability exemption in light of article 7 DSA.

However, codes of conduct can also create expectations on the turnaround times of notices, for example explicitly in the Code of Conduct on Countering Illegal Hate Speech+.[225] The question is whether voluntary codes of conduct are capable of affecting the level of diligence required in light of the liability exemption of article 6(1)(b) – is it possible for hosting providers to become liable after not abiding by a voluntary standard? Based on its monitoring reports, hosting providers have moved to responding to notices within 24 hours, as required by the Code of Conduct, but compliance varies. If expeditiousness by which hosting providers need to address content is interpreted as abiding by industry standards, then not abiding by those industry standards codified in the Code of Conduct may lead to liability for hosting illegal content, as it alters the duty of care for hosting providers on notices for illegal content. It is unclear whether these systemic obligations – 24 hours is a requirement generally, but not for notices specifically – can be extrapolated to individual liability cases. This remains to be seen in practice, but it is not impossible that courts will use codes of conduct to interpret what duty of care is expected from hosting providers, and use the norms set out therein to evaluate content moderation practices of hosting providers.

---

[221] Regulation 2022/2065, art. 45(1), see also rec. 103: "[…] While the implementation of codes of conduct should be measurable and subject to public oversight, this should not impair the voluntary nature of such codes and the freedom of interested parties to decide whether to participate. […]"

[222] Griffin (n 220) 175; Jacob van de Kerkhof, 'Jawboning Content Moderation from a European Perspective' in Charlotte van Oirsouw and others (eds), *European Yearbook of Constitutional Law: Constitutional law in the Digital Era*, vol 5 (1st edn, TMC Asser Press 2024).

[223] Regulation 2022/2065, rec. 103-104. It continues by stating that the mere fact of participation in a code of conduct does not create a presumption of compliance with this Regulation.

[224] Strengthened Code of Practice on Disinformation, Commitment 21.

[225] Code of conduct on Countering Illegal Hate Speech Online + https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online, 2.3.

III.iii Synthesis

Section I.i has given an overview of the current practice regarding intermediary liability exemption, and explained the ratio behind the exemption. The scope of intermediary liability exemption is explained broadly in CJEU case law. However, in *L'Oreal* and *YouTube,* the CJEU has been ambiguous in creating a duty of care for hosting providers. The DSA acknowledges that duty of care somewhat, but separates it from liability exemption. Intermediary liability exemption facilitates economic growth and prevents situations in which strict liability of hosting providers would cause harms to freedom of expression. However, Section II has illustrated that there are factors that require, and allow, a rethinking of *how* the standard of liability exemption is applied; although the phrasing of the liability exemption has remained relatively unchanged since the ECD, the landscape around that exemption has. Section III has explored provisions in the framework of the DSA and sectoral regulation that guide the interpretation of the intermediary liability exemption, and argues that negligent platforms can be excluded from the definition of hosting provider, and that duties of care created in the DSA and sectoral regulation can factually lower the threshold of awareness of illegal content thus triggering article 6(1)(b).

How does this argument reconcile with concerns for freedom of expression, prohibitions on general monitoring, and the Good Samaritan as outlined in I.ii? The suggestions above do not seek to exclude good faith hosting providers from liability exemption; it seeks to exclude bad faith hosting providers. The interpretation of the provisions addressed in this Section and their effect on the liability exemption should not be inconsistent with the prohibition on general monitoring obligations. They may lead hosting providers to monitor for identical or similar content as opposed to the exact illegality when the nature of illegality described in the notice requires such an intervention. They may also require hosting providers to monitor users that frequently upload manifestly illegal content. Neither option contravenes the prohibition on general monitoring obligations of article 8 as it is interpreted in CJEU case law. The Good Samaritan of article 7 protects hosting providers from being excluded from liability exemption for acquiring knowledge or awareness in good faith investigations on illegal content. Such may occur for example when investigating users under article 23, or fulfilling due diligence obligations under Codes of Conduct. However, article 7 does not protect against liability for not being

expeditious in removing such illegal content.[226] It merely ascertains that hosting providers remain neutral after undertaking measures to comply with requirements set by Union law or national law.

The most contingent point in this argumentation is the protection of freedom of expression: many fear that strict(er) liability can chill freedom of expression, as platforms will likely over-remove content to avoid liability.[227] Empirical data on this phenomenon can be divisive in the context of other regulations, for example in the case of *NetzDG*.[228] However, freedom of expression is not an absolute right, and needs to be balanced against other interests.[229] In its case law, the ECtHR has balanced the right to freedom of expression against people's right to private life,[230] but also included considerations whether someone's expressions, if they are considered hate speech, may undermine a cohesive society or suppress minorities.[231] This balancing act may be used to argue for better moderation practices by hosting providers, but it needs to be made carefully to prevent over-removal. The suggestions made above for more diligence by the hosting provider arguably stay within the limit set by the ECtHR for respecting the freedom of expression: the suggestions address gross negligence by the hosting provider, not small issues of non-compliance by *bona fide* hosting providers.

Finally, a note on proportionality. A narrow interpretation of liability exemption in line with additional requirements for diligence may be overly burdensome for smaller hosting providers. Although the requirements of article 16 DA also apply to micro- and smaller enterprises, many of the suggestions made above are also dependent on the position of the hosting provider. In case-by-case analyses, the capacity of smaller enterprises to meet the requirements laid out above can be weighed against the interests of aggrieved parties. This can be done in line with article 19 DSA, for example.

---

[226] Kuczerawy, 'The Good Samaritan That Wasn't' (n 100).

[227] E.g. Kuczerawy, 'The Power of Positive Thinking Special Issues' (n 7); Buiten, De Streel and Peitz (n 121) 161.

[228] E.g. 'Preventing "Torrents of Hate" or Stifling Free Expression Online?' (The Future of Free Speech 2024); Martin Eifert (ed), *Netzwerkdurchsetzungsgesetz in Der Bewährung: Juristische Evaluation Und Optimierungspotenzial* (1. Auflage, Nomos 2020); William Echikson and Olivia Knodt, 'Germany's NetzDG: A Key Test for Combatting Online Hate' (Centre for European Policy Studies 2018) 2018/09.

[229] E.g. European Court of Human Rights, 7 December 1976, *Handyside v the United Kingdom*, 5493/72, para 49.

[230] E.g. Court of Justice of the European Union, 13 May 2014, *Google Spain v Mario Gonzalez*, C-131/12; European Court of Human Rights, 16 June 2015, *Delfi v Estonia*, 64569/09, para 139; European Court of Human Rights, 25 February 2022, *Biancardi v Italy*, 77419/16, para 69.

[231] E.g. European Court of Human Rights, 9 May 2012, *Vejdeland et al v Sweden*, 1813/07, para 56-57.

Conclusion

The liability exemption for hosting providers in the ECD was adopted almost *verbatim* in articles 4-6 DSA. Conditional liability exemption provides an adequate but delicate balancing act between the interests of parties affected by illegal content, platform providers requiring legal certainty from being liable for illegal content, and the general public who have an interest in not being subjected to editorial control on the internet potentially constraining their freedom of expression. However, the standards developed around the application of the liability exemption stem from the early 2000s. In the meantime, platforms have developed into sophisticated networks where people socialise, shop, and take part in the democratic debate, all at the same time. Hosting providers have gained economically from hosting illegal content, and have developed sophisticated content moderation mechanisms. The current geopolitical climate puts effective content moderation under pressure. This article argues that in light of these developments we must rethink whether our interpretation of the liability exemption regime should evolve within the framework of the DSA.

The DSA introduces a range of due diligence obligations that encourage hosting providers to take more responsibility regarding their content moderation. Several provisions create implicit and explicit duties of care, both in the DSA or by sectoral regulation, and can factor into whether a service qualifies as a hosting provider, or whether it has met the threshold of awareness of illegal content. It is important to balance this interpretation on a case-by-case basis. The EU regulator had solid arguments for maintaining the conditional intermediary liability exemption; narrowing the application of the liability exemption requires balancing it with the prohibition on general monitoring *ex* article 8 DSA and the freedom of expression of internet users. However, in order to ensure a functioning right to freedom of expression, we need a functioning public sphere, which can be achieved through better combatting illegal content online and providing adequate remedies to aggrieved parties. Narrowing the scope of intermediary liability exemption may be a tool to achieve that, by incentivising platforms to take better care of their due diligence obligations and to better facilitate notice-and-takedown by users. Hosting providers have evolved significantly since the early 2000s; content moderation has evolved synchronously. Perhaps the intermediary liability exemption can evolve too.

Part II: The User

"The People have spoken.

Trump will be reinstated.

Vox Populi, Vox Dei"

Elon Musk after reinstating president Trump on X, 20 November 2022

Chapter IV: The Good, the Bad, the Compliant: Notice and Action Mechanisms under Article 16 DSA

Abstract: *The main remedy for users against illegal content online is a hosting provider's notice-and-action mechanism. Article 16 Digital Services Act (Regulation 2022/2065, 'DSA') requires hosting providers to make user-friendly and easily-accessible notice-and-action mechanisms available to anyone. Implementation of this article varies significantly per platform; this does not only raise compliance issues with article 16, but may distort the user's ability to effectively report illegal content, which can affect intermediary liability. This contribution presents a walkthrough of the notice-and-action mechanisms of AliExpress, TikTok, Twitch, and X, and reflects on compliance with article 16 and implications for the liability exemption of article 6 DSA. The walkthrough shows a paradox: the legal detail required for a successful notice creates a tension with article 16's requirement of user-friendliness. How platforms navigate this paradox is dependent on the nature of their services, but also affects their compliance with article 16, and even the framework of the liability exemption of article 6 DSA. Ultimately, the article argues that if the DSA is to be truly effective in the combat against illegal content, more guidance on article 16 and better enforcement of its requirements are needed.*

1. Introduction

In August 2023, X introduced a new affordance:[1] the 'Report EU illegal content'-button.[2] The button is distinct from the reporting affordances it is placed alongside, as an act of arguably performative compliance with the wishes of the EU regulator. X' new affordance is a response to the Digital Services Act (Regulation 2022/2065, 'DSA'),[3] the EU's landmark platform regulation.[4] The DSA requires all hosting service providers to provide a notice-and-action mechanism[5] to report illegal content[6] in article 16.[7] The rationale of this provision is that hosting providers have an important role in combatting illegal content, and should enable users to notify them of illegal content to take action.[8] Article 16 harmonises[9] requirements for these affordances, such as ease-of-access and user-friendliness, yet implementation looks differently across platforms. While X has introduced its new affordance very visibly, other platforms require users[10] to navigate through various interfaces before arriving at the article 16 notice-and-action mechanism. Previous research has indicated that different flagging modalities can influence users' flagging behaviour[11] and affect the legal outcome of reports submitted to the online platform.[12] Varying implementations of article 16 may be necessitated by the different natures of platform services, but non-compliant implementation can create a disconnect between the DSA's intention to combat illegal content on paper and its manifestation in practice.

---

[1] Generally: Taina Bucher and Anne Helmond, "The Affordances of Social Media Platforms" in Jean Burgess, Alice Marwick and Thomas Poell, *The SAGE Handbook of Social Media* (SAGE Publications Ltd 2018).

[2] Hura Anwar, 'X Is Rolling Out A New Button On The Web App To Report Illegal Content From The EU' (*Digital Information World*, 24 August 2023) <https://www.digitalinformationworld.com/2023/08/x-is-rolling-out-new-button-on-web-app.html> accessed 29 October 2025.

[3] Regulation 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).

[4] See generally: Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024); F Wilman, "De Digital Services Act (DSA): een belangrijke stap naar betere regulering van onlinedienstverlening" (2022) 28 Nederlands tijdschrift voor Europees Recht 220.

[5] For readability, this contribution uses 'mechanism' and 'affordance' in the context of notice-and-action mechanisms interchangably. Article 16 DSA phrases it as 'mechanism'.

[6] Regulation 2022/2065, art. 3(h).

[7] Regulation 2022/2065, art. 16.

[8] Regulation 2022/2065, rec. 50.

[9] Regulation 2022/2065, rec. 52.

[10] This contribution uses users and service recipients (the DSA's original phrasing) interchangeably for readability. This is also understood to include non-registered users of the platform.

[11] Ben Wagner and others, "Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <https://dl.acm.org/doi/10.1145/3351095.3372856> accessed October 18, 2023.

[12] Amélie Heldt, "Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports" (2019) 8 Internet Policy Review.

To examine how platforms have implemented article 16, this contribution examines **X, TikTok, AliExpress, and Twitch**. These platforms provide a variety of services ranging from social networking to online marketplaces. The walkthrough sheds light on compliance with article 16, but also on platform governance; design choices may affect how users interact with illegal content.[13] This contribution presents walkthroughs[14] of four notice-and-action mechanisms implementing article 16 DSA, offers a legal-doctrinal analysis using article 16 DSA, and questions associated consequences for liability exemption under article 6 DSA.

The article is structured as follows: Section 2 provides a background on notice-and-action mechanisms and article 16 DSA, as well as their role in intermediary liability exemption for hosting illegal content. Section 3 outlines the walkthrough methodology used to gather data and justifies the selection of the abovementioned platforms. Section 4 presents findings of the walkthrough, which are mapped against article 16 to assess compliance. Section 5 reflects on these findings by considering whether the implementation of these mechanisms affects the possibility for holding platforms liable for hosting illegal content. The article concludes that currently three out of four platforms have features that are non-compliant with article 16. Simultaneously, the article argues that article 16 creates a paradox between the user-friendliness of flagging mechanisms on the one hand and the legal accuracy required by article 16(3) and article 6(1) on the other. For meaningful user engagement in content moderation, platforms must better enable users to report illegal content; this can be guided by the European Commission and voluntary industry standards in line with article 44(1)(a).

2. Background

2.1 What are notice-and-action mechanisms?

Notice-and-action mechanisms are a ubiquitous element in content moderation. A notice-and-action mechanism marks a process, sometimes accompanied by a specific affordance, by which users submit reports on potentially harmful or illegal content ('flagging').[15] This marks a significant interaction between users and platforms. According to Crawford and Gillespie, flagging mechanisms fulfil at least two

[13] Generally Tarleton Gillespie, "Governance of and by Platforms" in Jean Burgess, Thomas Poell and Alice Marwick, *The SAGE Handbook of Social Media* (SAGE Publications 2017).

[14] Ben Light, Jean Burgess and Stefanie Duguay, "The Walkthrough Method: An Approach to the Study of Apps" (2018) 20 New Media & Society 881.

[15] Kate Crawford and Tarleton Gillespie, "What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint" (2016) 18 New Media & Society 410.

functions: (i) the first is that notice-and-action mechanisms provide an important practical mechanism to regulate vast amounts of content, and therefore fulfil a product-optimisation function; (ii) the second is that flags offer rhetorical legitimation to remove content. This shields platforms from further regulation, Crawford and Gillespie argue. "Flags are a practical and symbolic linchpin in maintaining a system of self-regulation",[16] as they maintain a narrative of self-regulation. Notice-and-action mechanisms allow platforms to 'hide' behind users reporting harmful or illegal content, and thus diminish the need for stringent regulation aimed at illegal content.

The mechanisms for reporting content have been explored from various perspectives in academic literature, for example from the perspective of the flagger,[17] from the perspective of the person whose content is flagged,[18] in the context of a specific medium[19] or in the context of a specific content type.[20] Legal perspectives on notice-and-action mechanisms have focused on establishing platform liability for hosting illegal content despite being notified,[21] and on compliance with (self)-regulatory instruments.[22] Wagner et al have reflected on notice-and-action mechanisms from a

[16] ibid 412.

[17] E.g. Wagner and others (n 10). Yunhee Shim and Shagun Jhaver, "Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms" (arXiv, September 13, 2024) <http://arxiv.org/abs/2409.08498> accessed November 27, 2024; Nora A Draper, "Distributed Intervention: Networked Content Moderation in Anonymous Mobile Spaces" (2019) 19 Feminist Media Studies 667, 668; Alice Qian Zhang, Kaitlin Montague and Shagun Jhaver, "Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts" (arXiv, September 11, 2024).

[18] Carolina Are, "Flagging as a Silencing Tool: Exploring the Relationship between de-Platforming of Sex and Online Abuse on Instagram and TikTok" [2024] New Media & Society 14614448241228544; Jessica T Feezell and others, "Who Gets Flagged? An Experiment on Censorship and Bias in Social Media Reporting" (2023) 56 PS: Political Science & Politics 222; Yao Lyu and others, "'I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.': A Study of Blind TikTokers' Content Moderation Experiences," *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2024).

[19] Lia Bozarth and others, "Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve This Process---A Case Study of COVID-19" (2023) 7 Proceedings of the ACM on Human-Computer Interaction 1; Paul B de Laat, "Coercion or Empowerment? Moderation of Content in Wikipedia as 'Essentially Contested' Bureaucratic Rules" (2012) 14 Ethics and Information Technology 123; Yubo Kou and Xinning Gui, "Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021).

[20] Michael Salter and Elly Hanson, "'I Need You All to Understand How Pervasive This Issue Is': User Efforts to Regulate Child Sexual Offending on Social Media" in Jane Bailey, Asher Flynn and Nicola Henry (eds), *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (Emerald Publishing Limited 2021); Kejsi Take and others, "Stoking the Flames: Understanding Escalation in an Online Harassment Community" (2024) 8 Proceedings of the ACM on Human-Computer Interaction 1; Shangyuan Wu, "What Motivates Audiences to Report Fake News?: Uncovering a Framework of Factors That Drive the Community Reporting of Fake News on Social Media" (2024) 12 Digital Journalism 790.

[21] Martin Husovec, *Injunctions against Intermediaries in the European Union: Accountable but Not Liable?* (1st edn, Cambridge University Press 2017); Aleksandra Kuczerawy, "Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative" (2015) 31 Computer Law & Security Review 46.

[22] Heldt (n 12); Eugénie Coche, "Privatised Enforcement and the Right to Freedom of Expression in a World Confronted with Terrorism Propaganda Online" (2018) 7 Internet Policy Review.

dark pattern perspective, suggesting that platforms steering users in flagging practices can be manipulative.[23] In legal literature, article 16 DSA has sparked some interest in studying notice-and-action mechanisms. Holznagel has signalled that article 16 compliance is an issue for TikTok.[24] In a more extensive study, Sekwenz et al interviewed legal specialists on the compliance of notice-and-action mechanisms of Facebook and TikTok with article 16 and the empowerment of users in that mechanism.[25] This article builds on previous literature by addressing specifically article 16 compliance, observing a different subset of platforms, and reflecting on article 16 from the perspective of liability exemption in article 6 DSA. The next subsections explore the development of article 16, its contents, and article 6 DSA.

2.2 Article 16 DSA

The role of notice-and-action mechanisms and article 16 DSA is rooted in conditional exemption from intermediary liability. Conditional liability exemption has been a cornerstone of platform regulation in the European Union since the e-Commerce Directive ('ECD').[26] The conditional liability exemption stipulates that once a hosting provider becomes has knowledge or becomes aware of illegal content, they must remove it without undue delay or risk liability for hosting said content.[27] If it is unaware of illegality, it is exempted from liability. Full liability for hosting illegal content is deemed undesirable because it would limit freedom of expression due to over-removal of content to avoid liability and hamper economic growth of hosting services due to compliance costs; conversely, full immunity would leave aggrieved parties with no remedies to ensure that illegal content is taken down. To resolve these tensions, the liability exemption of articles 12-14 ECD strike a balance between freedom of expression, economic interests of hosting services, and interests of parties aggrieved by illegal content.[28] Aggrieved parties may notify a hosting provider of illegal content, for which the hosting provider will be liable if it does not act

[23] Wagner and others (n 11).

[24] Daniel Holznagel, "Follow Me to Unregulated Waters!" [2024] Verfassungsblog <https://verfassungsblog.de/follow-me-to-unregulated-waters/> accessed September 23, 2024.

[25] Marie-Therese Sekwenz, Ben Wagner and Simon Parkin, "'It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!' – Evaluating Reporting Mechanisms under the Digital Services Act," *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2025).

[26] 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*, p. 1-16.

[27] Directive 2000/31, art. 14.

[28] Pieter Wolters and RaphaeI Gellert, "Towards a Better Notice and Action Mechanism in the DSA" (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 403, 405; Lilian Edwards, *Law, Policy, and the Internet* (Hart publishing 2019) 257.

expeditiously to remove that content. Conversely, the hosting provider is not liable for all content, as long as it maintains a neutral role in the content moderation process, meaning that its role is such that it does not raise awareness over content it hosts; it may however store, index, and recommend content through automated processes.[29] In the constellation of intermediary liability exemption, notice-and-action mechanisms are the primary avenue of triggering knowledge or awareness of illegal content; article 16 DSA therefore complements provisions on liability exemption, as is further explored in the next subsection.[30]

Notice-and-action mechanisms should provide a swift means of requesting action from internet intermediaries that is more efficient than judicial procedures while balancing the interests of hosting providers, users, and third parties.[31] Despite their crucial role in the liability of hosting providers, the process by which notice-and-action may take place was not harmonised in the e-Commerce Directive. Some member states have opted for a formalisation of the process, but most have waited for guidance by the Commission or self-regulation.[32] A notable example in national law is the German *Netzwerkdurchsetzungsgesetz*, which required that social media providers provide an 'easily recognisable, easily accessible, user-friendly and permanently available procedure' for submitting notices.[33] The EU regulator also made the non-binding requirement in Recommendation 2018/334 that hosting service providers have notice-and-action mechanisms that are "easy to access and use for all users", while also noting that "providers should remain flexible, for instance as regards the reporting format or technology to be used."[34] In sectoral regulation, requirements for notice-and-action mechanisms have also become increasingly common. For example, the Audiovisual Media Services Directive requires that video-sharing platforms establish operating and transparent and user-

[29] Court of Justice of the European Union, 22 June 2021, *Peterson v Google & Elsevier v Cyando*, C-682/18, para 114; Regulation 2022/2065, rec. 22; Folkert Wilman, "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" (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 317, 320–321.

[30] Folkert Wilman, Saulius Lukas Kalėda and Paul-John Loewenthal, *The EU Digital Services Act: A Commentary* (Oxford University Press 2024) 133.

[31] European Court of Human Rights, 16 June 2015, *Delfi v Estonia*, 64569/09, para 159. European Court of Human Rights, 2 February 2016, *MTE v Hungary*, 22947/13, para 81, 91. See also Aleksandra Kuczerawy, "From 'Notice and Takedown' to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression" in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020) 2.

[32] Aleksandra Kuczerawy, "The Power of Positive Thinking: Intermediary Liability as a Human Rights Issue" (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 226, 227.

[33] Netzwerkdurchsetzungsgesetz vom 1. September 2017 (BGBI. I S. 3352), para 3(1).

[34] Recommendation 2018/334 on measures to effectively T&Ckle illegal content online, *OJ L 63, 6.3.2018*, p. 50-61.

friendly mechanisms for users to report content.[35] The Regulation on addressing the dissemination of terrorist content online for reporting alleged terrorist content has a similar requirement for hosting service providers.[36] It is also implied in article 17 of the Copyright in the Digital Single Market Directive also implies that hosting providers must provide a mechanism for reporting copyright infringing content, especially when read in conjunction with the Commission's Guidance on that article.[37] The Commission suggests that member states align their implementation with Recommendation 2018/334 on illegal content.[38]

In this fragmented landscape article 16 DSA, despite applying without prejudice to the instruments cited above, provides a horizontal, harmonised set of requirements to which notice-and-action mechanisms must adhere, which can smoothen cross-border infringement procedures with unclarities by preventing unclarities on whether a notice was submitted.[39] Article 16 requires providers of hosting services to put mechanisms in place that allow any individual or entity to notify them of items of 'illegal information'.[40] These mechanisms must be easily accessible and user-friendly, and allow for the submission of notices by electronic means.[41] Article 16 leaves discretion for hosting providers to shape the mechanism along the needs of their respective service, as long as they adhere to the core principles of article 16, namely ease of access and user-friendliness.[42] The remainder of this subsection unpacks those principles.

---

[35] Directive 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive), *OJ L 303, 28.11.2018,* p. 69-92, art. 28b(3)(d).

[36] Regulation 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, *OJ L 172, 17.5.2021,* p. 79-109, art. 5(2)(b).

[37] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 20001/29/EC, art. 17.

[38] Communication from the Commission to the European Parliament and the Council Guidance on Article 17 of Directive 2019/790 on Copyright in the Digital Single Market; see also João Pedro Quintais and Sebastian Felix Schwemer, "The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?" (2022) 13 European Journal of Risk Regulation 191, 22–23.

[39] Ruth Janal, "Haftung Und Verantwortung Im Entwurf Des Digital Services Acts" [2021] Zeitschrift für Europäisches Privatrecht 227, 247.

[40] Regulation 2022/2065, art. 3(h) defines illegal content as: "any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law."

[41] Regulation 2022/2065, art. 16(1).

[42] As an example, Wikipedia was deemed compliant with article 16 in its audit by Holistic AI despite only providing an email address for submitting notices, thus providing the ultimate form-free mechanism. See on this: https://foundation.wikimedia.org/wiki/File:Wikipedia_DSA_Audit_Report_2023-24_Public.pdf

2.2.1 'Any individual or entity'

Firstly, the phrasing 'any individual or entity' indicates that *everyone* must be able to submit a notice not just registered service recipients.[43] This enables victims of illegal content and concerned parties to submit notices, despite not being registered users.[44] However, this does not require hosting providers to show information to everyone; if a hosting service is only accessible as a registered user, the consequence could be that only registered users will need to be able to use the notice-and-action mechanism.

2.2.2 'Easily accessible and user friendly'

Secondly, the notice-and-action mechanism must be easily-accessible and user-friendly. Raue identifies a number of features that factor into this.[45]

The first factor contributing to ease of access is that mechanisms must be clearly identifiable, meaning close to the information the user wishes to report, and equivalently findable as mechanisms for terms and conditions ('T&C'). The DSA requires this explicitly in recital 50: "[…] Such mechanisms should be […] *at least as easy to find* and use as notification mechanisms for content that violates the terms and conditions of the hosting provider. […]." (emphasis added)[46] Raue argues that this is intended to dissuade platforms from 'hiding' the report illegal content affordance in favour of T&C-reporting affordances.[47] The tendency to 'hide' illegal content affordance in favour of T&C affordances is due to the fact that it is generally easier and more attractive for platforms to moderate content on the basis of their terms and conditions. This allows them to interpret their own norms,[48] and remove content globally instead of geo-blocking it for the regions in which the content is considered illegal. As a result, most content moderation decisions are made on the basis of terms and conditions.[49] The distinction between non-compliance with terms and conditions and illegal content can be difficult to ascertain, and users can be under the

[43] Franz Hofmann and Benjamin Raue (eds), *Digital Services Act: Article-by-Article Commentary* (First edition, Nomos 2025) 328.

[44] Wolters and Gellert (n 28) 405–406.

[45] Hofmann and Raue (n 43) 333.

[46] Regulation 2022/2065, rec. 50.

[47] Hofmann and Raue (n 43) 330. In the context of NetzDG: Heldt (n 12); Wagner and others (n 11).

[48] Crawford and Gillespie (n 15) 412; João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, "How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications" (2023) 48 Computer Law & Security Review 105792.

[49] Rishabh Kaushal and others, "Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database," *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024). The equivalence of accessibility of article 16 notice-and-action mechanisms may affect transparency reporting obligations under article 15, 24, and 42 as well: if content is not flagged as illegal, it is likely not reported as such in transparency reports, skewing results. See Heldt (n 12).

impression that they flag content as 'illegal' while flagging it as incompliant with terms and conditions.[50] Because T&C-mechanisms do not necessarily match all the requirements of article 16(2), it is possible that users are not providing the platform with enough information, resulting in the platform dodging the legal consequences of notices submitted under article 16 such as exclusion from liability exemption.

Sekwenz et al note that article 25(1) DSA on manipulative design – so-called 'dark patterns'- can be used to support recital 50 on the aspect of location: providers of online platforms may not design their online interfaces in such a way that distorts their ability to make free and informed decisions – the same is true for notice-and-action mechanisms.[51] 'Hiding' a reporting affordance in favour of another severely affects users' ability to decide between the two affordances. A recent Dutch lower-level case provides some insight in how 'easily accessible' can be interpreted. In summary proceedings on whether options for recommender systems were 'easily accessible' under article 27(3) DSA, the Amsterdam District Court decided that 'hiding' the functionality to choose recommender systems under the logos of Instagram and Facebook respectively were not 'easily accessible', as nothing indicated to the user that those options were there.[52] Similarly, navigating a range of menus to change the recommender system (in the case of Facebook) was deemed not 'easily accessible'. In sum, although hosting providers enjoy a degree of discretion and these assessments need to be made on a case-by-case basis, it is clear that the notice-and-action mechanisms must be accessible without friction in a way that is easy to spot for the user.

The second point that can contribute to ease-of-access and user-friendliness is language. Language is a notably absent criterion in some respects in the DSA. It is unclear what language requirements article 16 poses to notice-and-action mechanisms, but the language in which the notice-and-action mechanism is available can affect its accessibility. Some guidance is provided in Article 11(3). Points of contact of hosting providers must be reachable in an official language of the EU member state of establishment, in addition to a language broadly understood by the largest possible number of EU citizens (generally English).[53] This must be respected by the competent authorities when issuing orders under article 9 and 10.[54]

---

[50] Heldt (n 12) 11.

[51] Sekwenz, Wagner and Parkin (n 25) 539.

[52] Rechtbank Amsterdam, 2 oktober 2025, *Bits of Freedom v Meta,* ECLI:NL:RBAMS:2025:7253, para 4.18 & 4.19.

[53] Regulation 2022/2065, art. 11(3); Hofmann and Raue (n 43) 252–253.

[54] Regulation 2022/2065, art. 9(2)(c) & 10(2)(c).

It is unclear whether this applies to notice-and-action mechanisms as well, but it can function as guidance. Raue suggests that reporting affordances must be available in the same language as the information that is reported, and that the phrasing must be accessible.[55] However, such variety of languages can be overly burdensome to smaller hosting providers. The balancing act made in article 11(3) may be more useful, namely that points of contact must be available in a language broadly understood by the largest possible number of EU citizens, i.e. English.

A third factor contributing to user-friendliness and accessibility is the complexity of the language used. Complex language can also affect the user-friendliness of the mechanism. Notice-and-action mechanisms must be contextually appropriate.[56] Notice-and-action mechanisms must be appropriate for the type of content hosted by the hosting provider, and appropriate for the platform's user base. For example, recital 89 stipulates that when a very large online platform or very large online search engine (VLOPSE) is used predominantly by minors it must make sure that their services are designed in a manner that is accessible to that specific audience.[57] In this regard, Wilman et al propose a 'context-specific' assessment, by which the user-friendliness of the mechanism must be assessed on the basis of the average type of recipient targeted by the service and the recipients that use the service.[58] The requirement of user-friendliness may create a paradox that is visible throughout the walkthrough. On the one hand, user-friendliness requires notice-and-action mechanisms to be phrased in a manner that is understandable to users.[59] On the other hand, submitting a notice that provides sufficient legal information to the hosting provider may require using legal jargon, which may contravene the user-friendliness. Striking a balance between legal accuracy and understandability is therefore paramount.

2.2.3 'By electronic means'

A final requirement of article 16(1) is that notice-and-action mechanisms must be electronic, as users must be able to submit notices in an exclusively electronic fashion.[60] This does not preclude the possibility of submitting notices through analogue means.

[55] Hofmann and Raue (n 43) 333.
[56] Janal (n 39) 252–253; Hofmann and Raue (n 43) 330.
[57] Regulation 2022/2065, rec. 89.
[58] Wilman, Kalėda and Loewenthal (n 30) 136.
[59] Sekwenz, Wagner and Parkin (n 25).
[60] Regulation 2022/2065, art. 16(1).

2.2.4 'Sufficiently precise and adequately substantiated notices'

As the purpose of these mechanisms is to alert hosting providers of illegal content and create legally enforceable takedown, users must be able to submit sufficiently precise and adequately substantiated notices to enable the hosting provider to ascertain itself that removing content is not incompatible with freedom of expression without a detailed legal examination.[61] This criterion is iterated in recital 53, but stems from *Youtube/Cyando,* in which the Court of Justice of the European Union decided on the requirements of a notice that would render a platform aware of illegal content and thus preclude liability exemption.[62]

Notices must, in order to trigger the notice-and-action mechanism that could lead to intermediary liability, contain the following elements: (i) a sufficiently substantiated explanation of the reasons why the individual or entity alleges the information in question to be illegal content; (ii) a clear indication of the exact electronic location of that information; (iii) the name and email address of the individual submitting the notice; and (iv) a statement confirming bona fide belief that the notice is accurate and complete.[63] Wilman et al note that the first two elements can suffice to submit a notice capable of triggering intermediary liability in relation to article 16(3).

Article 16(2)(i) requires that a notice contains a sufficiently substantiated explanation. This must explain why the content is illegal, but also the nature of the illegality. To facilitate that, notice-and-action mechanisms must not constrain the user in providing detailed information. A notice-and-action affordance with an overly limitative word- or character count would be incompliant, for example, as the user may not be able to explain why the content is illegal. Another feature that contributes to this requirement is the user's ability to upload evidence in their process of reporting.[64] This can expand their possibility to describe the illegality of the content, and present evidence. This could be particularly relevant in the case of intellectual property infringement, where the user might be required to demonstrate the existence of an IP-right. It is here that the paradox mentioned above becomes most visible; a sufficient amount of detail is required to explain why the content is

[61] Regulation 2022/2065, rec. 53.

[62] Court of Justice of the European Union, 22 June 2021, *YouTube/Cyando,* C-682/18 and C-683/18, para 116, also Opinion of Advocate General Saugmandsgaard Oe, para 190; Court of Justice of the European Union, 26 April 2022, *Commission v Poland,* C-401/19, para 91 See also Janal (n 39) 247. On the adoption of the case law of the Court of Justice of the European Union in the DSA regarding intermediary liability exemption, see Ronan Riordan, "A Case Study of Judicial-Legislative Interactions via the Lens of the DSA's Host Liability Rules" (2025) 2025 10 European Papers - A Journal on Law and Integration 259.

[63] Regulation 2022/2065, art. 16(2).

[64] Janal (n 39) 252.

illegal, but it can be overly burdensome for the average user to explain this exhaustively.[65]

Article 16(2)(ii) requires a precise indication of the location of the illegal content. Usually, this is specified using uniform resource locators. Oftentimes this requirement is met by a prefilled URL in cases where the notice-and-action affordance is located next to the content that is being reported on.

Article 16(2)(iii) requires that the notice contains the identity of the user submitting it, with the exception of reports relating to Child Sexual Abuse Material ('CSAM').[66] The requirement for identification raises the question whether anonymous notices on other types of content are allowed and whether they must be facilitated. Recital 50 hints that other notices may also be submitted anonymously.[67] This is also supported by the requirement that 'anyone' must be able to submit notices; identification can discourage people to report illegal content. Conversely, anonymity has its downsides. Without identification, there is no possibility for feedback or follow-up with the notifying party, which may hamper requirements posed in article 16(5).[68] Similarly, without the identity of the person submitting the notice, the illegality of some types of content cannot be assessed.[69] Finally, in cases of abuse,[70] it may be more difficult to suspend the flagging party in light of the requirements of article 23(2) requiring the online platform to suspend notifying privileges for abusive users. While this may be overcome to some extent with the use of an anonymous email-address or with anonymous user identification software,[71] but anonymous reports still raise questions on the degree to which they enable follow-up for evidence-gathering or notification. As it stands, the hosting provider must allow the notifying party to identify itself, but it is not a requirement for the submission of a notice.

---

[65] Wilman, Kalėda and Loewenthal (n 30) 137.

[66] Regulation 2022/2065, art. 16(2)(c). See also Wolters and Gellert (n 28) 414. CSAM is defined by any offences under art. 3-7 of Directive 2011/93/EU, see also rec. 53.

[67] Regulation 2022/2065, rec. 50: "[…] The notification mechanism *should allow, but not require*, the identification of the individual or the entity submitting a notice. For some types of items of information notified, the identity of the individual or the entity submitting a notice might be necessary to determine whether the information in question constitutes illegal content, as alleged. [referring to copyright] […]." [emphasis added].

[68] Regulation 2022/2065, art. 16(5).

[69] Regulation 2022/2065, rec. 50.

[70] See on abuse e.g. Jennifer M Urban, Joe Karaganis and Brianna L Schofield, "Notice and Takedown in Everyday Practice" (University of California, Berkeley 2017) 127–129.

[71] Oleksandr Korchenko, Anna Korchenko and Ivan Azarov, "Analysis of Modern Solutions for the Identification of Anonymous Users" in Stanislav Dovgyi and others (eds), *Applied Innovations in Information and Communication Technology* (Springer Nature Switzerland 2025).

Article 16(2)(iv) requires that the user indicate their good faith belief that the notice is accurate and complete. This measure is intended to prevent abuse;[72] however, if the notice is submitted anonymously, the statement is essentially symbolic. The legal consequences of submitting a bona fide belief statement are also unknown.

Assuming that a notice is sufficiently substantiated, it can trigger the obligation for takedown if the content is illegal. For hosting providers, this is regulated in the conditional intermediary liability provision of article 6 discussed in the next subsection.

2.3 Article 6 DSA

The previous subsection hinted at the possibility of notice-and-action mechanisms affecting liability exemption. Liability for illegal content is not harmonised in the DSA,[73] but liability exemption is. Generally, when hosting providers perform a neutral role in content moderation and do not have knowledge of illegal content or act expeditiously upon learning of illegal content, they are exempt from liability for hosting that content.[74] If the hosting service has adequately facilitated the provision of notices in this notice-and-action mechanism, a successful notice leads to actual knowledge of illegal content.[75] Upon receipt of a notice, a hosting provider must process the notice in a timely, diligent, non-arbitrary, and objective manner.[76] If it fails to remove the illegal content without undue delay, this leads to exclusion from the intermediary liability hosting exemption. Notice-and-action mechanisms play a crucial role in this framework: they are the primary, but not only,[77] avenue for users to 'create' 'actual knowledge' about illegal content with the hosting provider, triggering the duty to take action under article 6.[78] This framework is explicitly acknowledged in article 16(3), linking the notices submitted through the mechanisms discussed above to the liability exemption.[79] It is therefore paramount that this avenue is functional and well-designed, otherwise users may not be empowered to trigger liability and enforce takedown of illegal content.

[72] Wilman, Kalėda and Loewenthal (n 30) 138.
[73] Regulation 2022/2065, art. 3(h).
[74] Regulation 2022/2065, art. 6(1).
[75] Regulation 2022/2065, art. 16(3).
[76] Regulation 2022/2065, art. 16(6).
[77] See Kammergericht Berlin, 25 August 2025, 10 W 70/25, deciding that users need not use article 16 notice-and-action mechanisms to exercise their rights.
[78] Wolters and Gellert (n 28).
[79] Regulation 2022/2065, art. 16(3).

Further it is important to note that the distinction between illegal content, as defined in article 3(h), and content that is non-compliant with the terms and conditions, in notice-and-action mechanisms, is crucial. As noted, platforms must be notified with an adequately substantiated and sufficiently precise notice. The notice-and-action framework must facilitate the provision of such information for illegal content, otherwise it is impossible to hold hosting providers liable within the framework of article 6. The information necessary cannot always be provided in T&C mechanisms; it is not required for removal requests under terms and conditions, as those fall under the full discretion of the hosting provider. This *de facto* incapacitates the user to fulfil the requirements of article 6 and creates an obligation for the platform to remove that content. Not only does it affect the position with regards to intermediary liability of the hosting provider, it also affects the notification requirements that hosting providers have with regards to the submitter of the notice in relation to articles 16(4) and 16(5); the hosting provider is under no obligation to alert the reporter of a T&C notice of what decision was taken as a result of that notice. It is therefore crucial to make a distinction between the mechanisms that are designed for illegal content and T&C-mechanisms, since the legal outcome can be very different.

2.4 Summary

Article 16 harmonises requirements for notice-and-action mechanisms. The DSA requires that *anyone* must be able to submit notices; notices on CSAM can be anonymous, and on other content they may be anonymous but may trouble requirements under article 23(2). The requirements for ease-of-access and user-friendliness are undefined in the DSA, and require interpretation on a case-by-case basis. Mechanisms must be located near the information for which notices can be submitted, and be similarly findable as reporting for T&C compliance. Mechanisms must also be contextually appropriate, both from a language and an understandability perspective. Notice-and-action mechanisms are inseparably linked to hosting liability, and must therefore allow for submitting sufficient information on the illegal content along the requirements of article 16(2). As the primary avenue for requesting hosting providers to take down illegal content in the context of article 6, it is paramount that they are well-designed and comply with article 16. Before moving to the walkthrough assessing these requirements, the next Section outlines the methodology for conducting the walkthrough.

3. Methodology

3.1 The Walkthrough Method

To facilitate the systematic and cross-platform comparison of interfaces, this article draws on the walkthrough method. Light et al developed the walkthrough method for studying technological systems to "[recognise] discursive and symbolic representations", meaning to study how meaning is created or steered through technological systems by engaging with those systems.[80] The method is used to chart social media and internet applications,[81] but also to evaluate compliance with legal requirements.[82] Walkthroughs allow for direct engagement with an application's interface to "examine its technological mechanisms and embedded cultural references to understand how it guides users and shapes their experiences."[83] This method is therefore suitable for the specific requirements of article 16, particularly the user-friendliness and ease-of-access.[84]

The walkthrough method has limitations: Duguay and Gold-Apel note that the walkthrough method suffers from (i) the personalised and adaptive nature of social media, and (ii) the personal bias of the researcher.[85] The nature of this study mitigates those limitations. The personalised and adaptive nature of social media does not play a significant factor in flagging mechanisms studied, which are standardised reporting tools to adhere to legal requirements, and therefore not adaptive to the user.[86] However, mechanisms are adapted over time,[87] which is why the walkthrough is limited to a specific timeframe. Further, the personal bias of the researcher can be mitigated somewhat by referring to findings in auditing reports

[80] Light, Burgess and Duguay (n 14) 884.

[81] E.g. Mathias Decuypere, "The Topologies of Data Practices: A Methodological Introduction" (2021) 10 Journal of New Approaches in Educational Research 67; Giovanna Mascheroni and Donell Holloway, "Hybrid Methods for Hybrid Play: A Research Toolkit" in Giovanna Mascheroni and Donell Holloway (eds), *The Internet of Toys: Practices, Affordances and the Political Economy of Children's Smart Play* (Springer International Publishing 2019).

[82] See Taylor Annabell, Laura Aade and Catalina Goanta, "(Un)Disclosed Brand Partnerships: How Platform Policies and Interfaces Shape Commercial Content for Influencers" (2024) 13 Internet Policy Review.

[83] Light, Burgess and Duguay (n 14) 882.

[84] Regulation 2022/2065, art. 16(1).

[85] Stefanie Duguay and Hannah Gold-Apel, "Stumbling Blocks and Alternative Paths: Reconsidering the Walkthrough Method for Analyzing Apps" (2023) 9 Social Media + Society 20563051231158822.

[86] Sekwenz et al report that the report illegal content functionality of Facebook changes according to jurisdiction, offering a varying amount of legal grounds to report content on across France, Austria and Germany. In the notice-and-action mechanisms examined in this contribution, namely those of AliExpress, TikTok, Twitch, and X, this does not appear to be the case. See Sekwenz, Wagner and Parkin (n 25) 539.

[87] For example, KPMG mentions in AliExpress' DSA Audit report that they have made significant changes to their notice-and-action mechanism in 2024, accessible via https://www.aliexpress.com/p/transparencycenter/auditReport.html, p. 28.

(available for 3 out of 4 platforms)[88] and contextualising the findings against empirical evidence of the Transparency Database holding *inter alia* statements of reasons related to notice-and-action mechanisms.[89] The audit reports were consulted after the walkthrough to prevent bias. Reflections on the auditing reports are integrated throughout the walkthrough to evaluate findings, but primarily reflected on specifically in Section 4.5.

3.2 Case selection and environment of expected use

This contribution studies notice-and-action mechanisms of X, AliExpress, Twitch, and TikTok. These platforms are selected for a descriptive case study based on diverse cases, "aimed at capturing the diversity of [the] subject."[90] Although this can complicate making any causal interferences or generalisable statements, but it does, in this case, show compliance across four different platforms. This section describes the chosen platforms and provides an insight in their 'environment of expected use.'[91] The environment of expected use as Light et al describe it is the socio-economic and cultural context around the technical walkthrough in which the affordance is interrogated. This context, consisting of aspects such as vision, operating model, and governance, is used to accurately describe the usage and experience of an application. Since the compliance with article 16 is largely determined by the results of the technical walkthrough and not the environment of expected use, this subsection suffices with a brief overview.

X hosts short-text contributions, longer text contributions, and audio-visual media. X targets a universal user base, but following its takeover by Elon Musk it is increasingly populated by right-wing users as left-wing users flock to other platforms.[92] X generates revenue through selling advertisements, selling data licenses, and offering a paid subscription service. It also has a unique relationship with DSA compliance.[93] Musk has been a verbal opponent of the Digital Services Act. X has publicly resisted DSA compliance and withdrawn from earlier co-regulatory

[88] Yearly reports submitted by all Very Large Online Platforms and Very Large Online Search Engines (Regulation 2022/2065, art. 33(1)) reporting on an independent audit conducted on their services' compliance with Chapter III (Regulation 2022/2065, art. 37). Interestingly, most auditing reports also use walkthroughs as a method for checking compliance with article 16.

[89] Regulation 2022/2065, art. 24(5); see also Kaushal and others (n 49).

[90] John Gerring, *Case Study Research: Principles and Practices* (Second edition, Cambridge University Press 2017) 58.

[91] Light, Burgess and Duguay (n 14) 889–891.

[92] Elisa Shearer Matsa Sarah Naseer, Jacob Liedke and Katerina Eva, "X Users' Experiences with News" (*Pew Research Center*, June 12, 2024) <https://www.pewresearch.org/journalism/2024/06/12/x-users-experiences-with-news/> accessed October 13, 2025.

[93] For an overview of the Commission's proceedings against X: Matteo Fabbri, "The Role of Requests for Information in Governing Digital Platforms Under the Digital Services Act: The Case of X" (2025) 6 Journalism and Media 41.

efforts targeting content moderation. As a result, his platform is targeted both nationally and internationally in private and public enforcement of the DSA.

TikTok offers a short-form video-based platform. Although it facilitates a universal userbase with 1.6 billion active monthly users, it has an overall younger audience than the other platforms.[94] It generates revenue through advertisements, brokering data, and in-app purchases and e-Commerce affordances. TikTok is currently under investigation by the European Commission for its recommender systems, and has already been found in breach for its ad repository.[95]

AliExpress is an online marketplace that facilitates e-Commerce for predominantly Chinese sellers. It has a user base of about 150 million active monthly users, predominantly consumers between 18 and 65. It is expected that, building on the different nature of this platform compared to the previous two, it has a different approach to content moderation. This may target predominantly IP infringement, dangerous products, and user reviews. AliExpress generates revenue through commissions and fees for facilitating transactions, advertising, and data brokerage. AliExpress has already been found in breach of the DSA, with regards to its risk mitigation, notice-and-action mechanism, complaint handling system, advertiser transparency, trader traceability, and access to public data.[96] This has resulted in binding commitments with the EU Commission, the effects of which should be visible in this walkthrough.[97]

Finally, Twitch is a livestream video-based platform thus differing in service from the previous services. Twitch' user base is predominantly males under the age of 35. It generates revenue from advertisements, offering paid subscriptions, and in-app e-Commerce affordances. Twitch is not a Very Large Online Platform or Very Large Online Search Engine ('VLOPSE'),[98] with only 28.4 million monthly active users. This makes it impossible to reference to audit reports, as these are only required for VLOPSEs. Additionally, implementation of the DSA was required earlier for

[94] Generally: Radu M Meza, Andreea-Alina Mogoș and George Prundaru, "Idols of Promotion and Authenticity on TikTok" (2023) 11 Media and Communication.

[95] 'Commission preliminary finds TikTok's Ad Repository Breach Digital Services Act, (*European Commission,* 2025), <https://digital-strategy.ec.europa.eu/en/news/commission-preliminarily-finds-tiktoks-ad-repository-breach-digital-services-act> accessed 29 October 2025.

[96] 'Commission opens formal proceedings against AliExpress under the Digital Services Act', (*European Commission,* 2025), <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485> accessed 29 October 2025.

[97] 'Commission makes AliExpress commitmens under the Digital Services Act binding' (*European Commission,* 2025), <https://digital-strategy.ec.europa.eu/en/news/commission-makes-aliexpress-commitments-under-digital-services-act-binding> accessed 29 October 2025.

[98] Regulation 2022/2065, art. 33(1).

VLOPSEs than for other platforms;[99] therefore it is possible that notice-and-action mechanisms on X, AliExpress, and TikTok may be further developed than Twitch's.

The platforms are chosen because they can be considered important platforms in the European landscape, and because they have stark differences with regards to the services that they provide and their relation with the Digital Services Act. Each platform has individually received media attention for the impact their services make in society and their complicated relation with illegal content. Furthermore, where X, AliExpress, and TikTok have already been targeted by DSA enforcement actions by the European Commission and national authorities. X and AliExpress have already been under investigation for article 16 compliance, TikTok has not. Despite their differences the platforms overlap to some extent on the amount of notices they report having received under article 16, ranging from 300k to 350k, despite varying sizes in user base. This is also reflected on under Section 4.5. Twitch is included because platforms below the VLOPSE-barrier can indicate gaps in DSA enforcement, as is the case for Telegram[100] and Kick,[101] for example. Twitch has previously caused unrest as its services were used to livestream mass-shootings, underscoring the need for a functioning notice-and-action mechanism. It is worthwhile to assess how this position affects compliance with article 16 in comparison to VLOPSEs. The diversity of these hosting services can indicate the variety in uptake of article 16 requirements across a range of actors that have different stances towards DSA enforcement.

3.3 Approach

To conduct the technical walkthrough, a new account is registered for each platform and used to submit notices and mimic everyday 'flagging' activities. In addition, the ability to flag as a non-registered user is examined. Notice-and-action mechanisms are observed in a step-by-step manner. These affordances can, after documentation,

[99] Regulation 2022/2065, art. 93(2) only requires implementation since 17 February 2024. For VLOPSEs this was four months after their designation as VLOPSE (art. 92).

[100] "Telegram Still Doesn't Meet Large Platform Requirements under DSA" (*Yahoo*, 2024) <https://sg.news.yahoo.com/telegram-still-doesnt-meet-large-095311777.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGA_0t6lA009T3PlmMHDv1K228kVtDkXBlkuappqM2CuEPAlK-YLwk0wp5XpT96_vaubJGnEV1UPgbwFNEBQsyAQ7o2LyJE8V8Y5na55OlvFvRvSUQEKMWwxHUZzHcvzrMzAVkyJforkxdgsxHWdAI6LoHmo7upC6Icn_8fnfQMi> accessed November 1, 2024.

[101] Euractiv, "How the Tragic Death of a French Streamer Exposes Flaws in DSA Enforcement" (*Euractiv*, 2025) <https://www.euractiv.com/news/how-the-tragic-death-of-a-french-streamer-exposes-flaws-in-dsa-enforcement/> accessed October 8, 2025.

be held to the standards of article 16 DSA. The walkthrough was conducted over March to June 2025, with all affordances being up to date as recent as June 2025.[102]

The walkthrough uses the latest iOS-version of each application and the web-based interface. Since all platforms are accessible through mobile applications and web-based interfaces, and notice-and-action mechanisms need to function for all means of access. To standardise what content is reported, the study focuses on reporting illegal hate speech. Although hate speech can be considered a grey area speech type, there is increasing consensus in the European Union on its illegality.[103] The illegality of hate speech is explicitly acknowledged in the Framework Decision on combating certain forms and expressions of racism and xenophobia[104] and also developed in European Court of Human Rights case law.[105] The additional benefit of using this specific content type is that it is targeted by the Code of Conduct on Illegal Hate Speech,[106] recently adopted under the DSA.[107] Although these codes of conduct are voluntary instruments, non-participation is seen as an indicator of non-compliance with the DSA, meaning that particularly VLOPSEs (all studied platforms except Twitch) are incentivised to adhere of the Code of Conduct.[108] In addition to hate speech reporting functions, the walkthrough observes the functionality of reporting CSAM content due to the DSA's explicit requirement for an anonymous reporting functionality.[109]

4. Walkthrough and Article 16 Compliance

This section shows the results of the technical walkthrough of the notice-and-action mechanisms of AliExpress, TikTok, Twitch, and X through the lens of article 16. Full documentation of the walkthrough can be found in an open-access repository.[110] The findings of the walkthrough are discussed along the lines of the criteria outlined above: (1) notice-and-action mechanisms must be for 'any individual or entity', (2)

[102] Jacob van de Kerkhof, 'Walkthrough article 16 DSA Notice and action mechanisms', *DataverseNL* <https://doi.org/10.34894/VQV3VW> accessed 29 October 2025.

[103] Natalie Alkiviadou, "Hate Speech on Social Media Networks: Towards a Regulatory Framework?" (2019) 28 Information & Communications Technology Law 19.

[104] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, *OJ L 328, 6.12.2008*, p. 55–58

[105] E.g. European Court of Human Rights, 15 October 2015, *Perincek v Switzerland*, 27510/08; European Court of Human Rights, 16 July 2009, *Feret v Belgium*, 15616/07; European Court of Human Rights, 23 September 1994, *Jersild v Denmark*, 15890/89.

[106] In the meantime this has been rebranded to the Code of Conduct+. Out of the platforms studied, it was signed by TikTok, X, and Twitch.

[107] Regulation 2022/2065, art. 45.

[108] Regulation 2022/2065, rec. 104.

[109] Regulation 2022/2065, art. 16(2)(c).

[110] Van de Kerkhof (n 102).

they must be easily accessible in terms of location, (3) they must be user-friendly (4) they must facilitate submission of an adequately substantiated and sufficiently precise notice along the requirements of article 16(2), and (5) the notices must be electronic. The technical walkthrough indicates that the latter point is not of interest from a compliance perspective, and therefore the walkthrough addresses the former four points.

4.1 'Any individual or entity'

Article 16 requires that notices can be submitted by anyone, meaning without logging in, and raises the question whether anonymous reporting is also a requirement. The walkthrough has attempted to report hate speech both logged in and without logging in. It paid particular attention to the possibility of submitting notices anonymously for hate speech *and* for CSAM, as CSAM has an additional legal requirement for anonymity not present for other types of content. All studied platforms allow for registering an account without disclosing personal identity, meaning that, if no further information is required in the notice-and-action mechanism, a notice can be submitted anonymously when logged-in to the account.

In assessing whether 'any individual or entity' can notify the hosting service of the presence of illegal information, it is important to make a distinction between the iOS interface and the web interface. TikTok, Twitch, and X do not allow accessing content on the iOS application without being a logged-in service recipient. Although notices cannot be submitted, there is also no content to report about. As discussed above, hosting providers are under no obligation to make content available to everyone, and therefore it is not inconsistent with article 16 to not have a notice-and-action mechanism when not displaying content. Conversely, AliExpress allows users to access the iOS application and submit notices without logging in, but through a different portal.[111] This portal encourages users to log in before reporting illegal content (Figure 1). The portal specifically allows submitting notices on a variety of reasons including CSAM[112] and life-threatening criminal offenses.[113]

[111] This is also acknowledged in their transparency report, see p.7 https://www.aliexpress.com/p/transparencycenter/transparencyReport.html

[112] 'Content reported involves one of the offences referred to in Article 3 to 7 of Directive 2011/93/EU (including sexual abuse and sexual exploitation of children and child pornography).'

[113] 'Criminal offense involving threat to life or safety or persons pursuant to Article 18 of Digital Service Act'

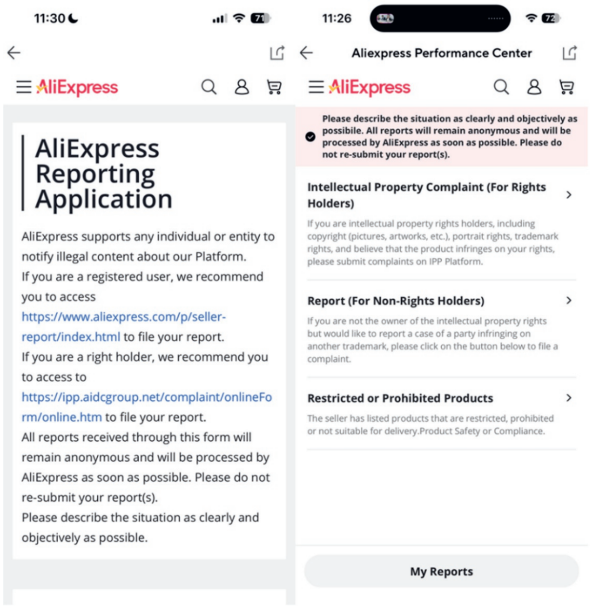


Figure 1: AliExpress mobile not logged-in vs logged in

On the web-based interface, X, TikTok,[114] and AliExpress offer the possibility for *anyone* to submit notices on illegal content, irrespective of whether they are a registered user. AliExpress offers the same functionality to registered users as it does to non-registered users, but it encourages registered users to use the mechanism for registered users. This presumably speaks to the requirements of article 23(2). X does not allow users that are not logged-in to access the T&C notice-and-action mechanism, only allowing the 'report EU illegal content' option. When logged in, the user is provided with more options (Figure 2).

[114] Over the course of the review process of this contribution, TikTok has altered the accessibility to the web-based interface for users that are not logged-in. This is not reflected in the data as presented in the repository. This does not significantly alter the findings of this contribution, however.



Figure 2: Reporting content on X logged in vs not logged in

TikTok's reporting interface on the web (figure 3) is unaffected by a user being logged-in or not. The same is true for Twitch. In sum, the requirement for 'anyone' to submit notices is met across the platforms studied. The question of anonymity, which is also related to whether 'anyone' can submit notices (see Section 2.2.4), is discussed further under Section 4.4.

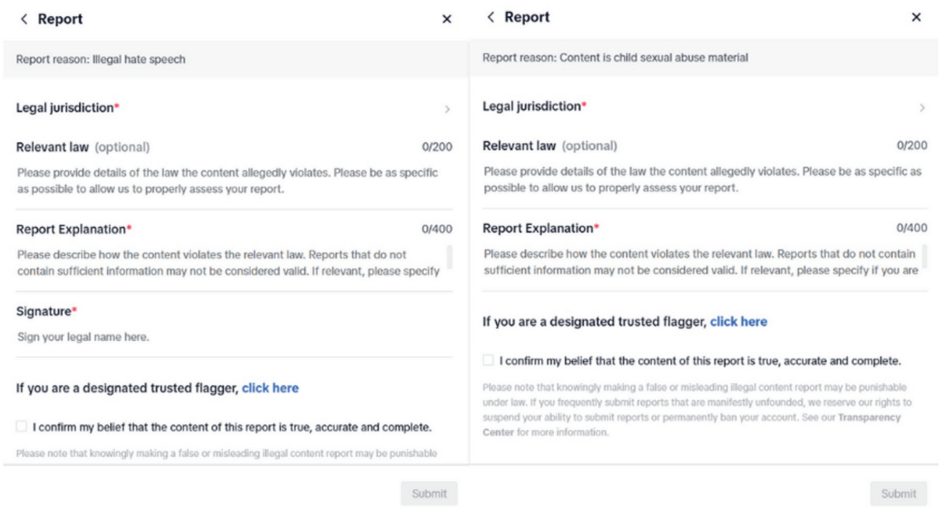


Figure 3: TikTok removing 'legal name' requirement for CSAM (right) as opposed to hate speech (left)

4.2 'Easily accessible'

To be easily accessible, the notice-and-action mechanism must be located near the information the user is reporting on. For all platforms, accessing the notice-and-action mechanism requires at least two interactions, often through an ellipsis

menu.[115] Table 1 displays the amount of interactions required before reaching the article 16 notice-and-action mechanism.

Platform	**Interactions on web**	**Interactions on mobile**
AliExpress products[116]	2: Significant scrolling, report item button, selecting the article 16 mechanism	2: Significant scrolling, report item button, selecting the article 16 mechanism
AliExpress reviews	2: ellipsis button, report review button	2: ellipsis button, report review button
TikTok	3: ellipsis button, report content, report illegal content	3: long press/share button, report, report illegal content
Twitch[117]	2: ellipsis button, report livestream	2: ellipsis button, report livestream
X	2: ellipsis button, report EU illegal content	2: ellipsis button, report EU illegal content

Table 1: How many interactions are required before reaching the notice-and-action affordance?

The number of interactions is relevant as they may introduce friction – defined as "points of difficulty encountered during users' interaction with a technology"[118] – to reporting on illegal content, and complicates the reporting functionality.

For products listed on AliExpress, the 'report item' button is placed next to the product description (Figure 4). This seems close to the information that is reported on, but in practice it requires significant scrolling for the user in both the iOS and web interface to reach that affordance. The product can even be purchased without ever seeing a 'report item' button. This is aggravated by the fact that there is no indication for the user that the reporting affordance will be located somewhere on the page; the same is true for the reporting affordance located all the way at the footer of the webpage. It can be questioned whether this is easily-accessible. AliExpress' audit report does not address this matter as it is currently under

[115] Eric S Lee and Darrell R Raymund, "Menu Driven Systems" in Allen Kent and James G Williams, *The Encyclopedia of Microcomputers*, vol 11 (1993). Also depending on the rotation: 'meatball menu' or 'kebab menu'.

[116] AliExpress notice-and-action mechanism does not clarify whether it considers notices submitted as article 16 notices.

[117] In Twitch' notice-and-action mechanism it is unclear whether content is reported as illegal or as non-compliant with terms and conditions. This is by design, according to their transparency report (accessible via: https://safety.twitch.tv/s/article/Twitch-DSA-Transparency-Report-February-2025?language=en_US). The transparency report clarifies that all reports are treated as an article 16 notice.

[118] Anna L Cox and others, "Design Frictions for Mindful Interactions: The Case for Microboundaries," *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Association for Computing Machinery 2016); Thomas Mejtoft, Sarah Hale and Ulrik Söderström, "Design Friction," *Proceedings of the 31st European Conference on Cognitive Ergonomics* (Association for Computing Machinery 2019).

investigation for article 16(1) compliance.[119] The report does note that the avenues of reporting illegal content were not located clearly enough across the many affordances AliExpress offers.[120]

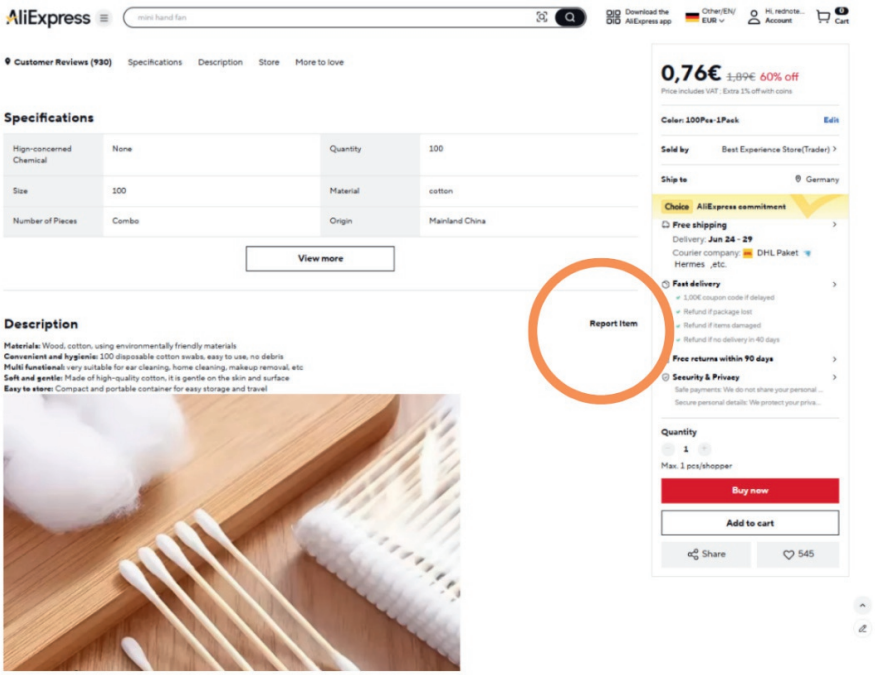


Figure 4: AliExpress Report Item

AliExpress has a different reporting mechanism for reviews (Figure 5); those are accessed through an elliptical menu located next to the review that is reported on, and therefore easily accessible.

[119] 'Commission opens formal proceedings against AliExpress under the Digital Services Act', (*European Commission*, 2025) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485>, accessed 29 October 2025.

[120] 'Audit Report', (*AliExpress*, 2025), <https://www.aliexpress.com/p/transparencycenter/auditReport.html>, p. 25-27, accessed 29 October 2025.

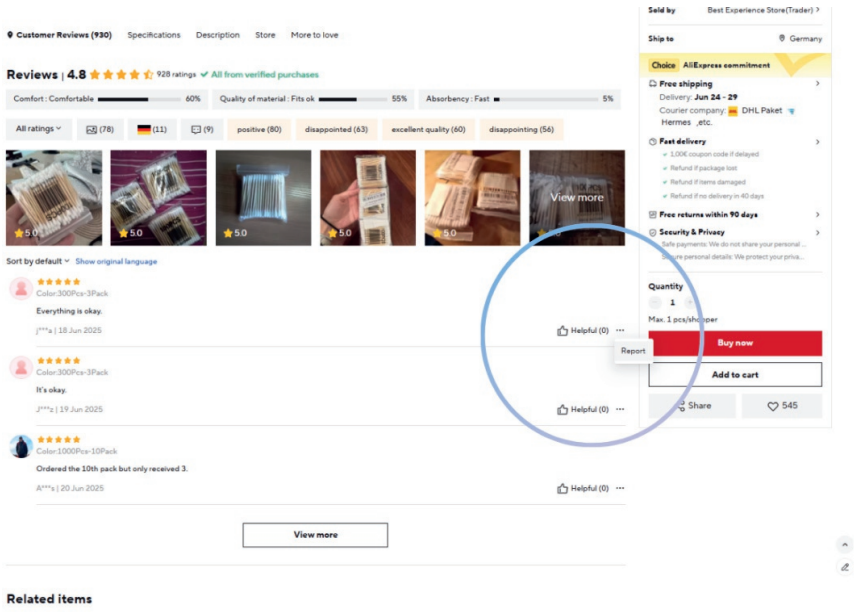


Figure 5: AliExpress Report Review

Twitch and X use the ellipsis menu next to the information across mobile and desktop interfaces to access the notice-and-action mechanism. On X, this menu is always located next to the information reported on (see for example Figure 2). However, Community Notes, arguably user-generated information, cannot be reported.[121] Community Notes have an exceptional position because they are a content moderation tool, user-generated information, and subject to platform editorial control, all at the same time.[122] In the editorial phase, before the Community Note is visible to all users, users in the Community-note group can report notes using a specific affordance. Once published, this is no longer possible. It is unclear whether this raises a compliance issue with article 16, since the content is not publicly visible. Twitch also has a notable feature regarding the location of article 16 mechanisms: reporting the chat next to a livestream (Figure 6). It is not possible to flag individual messages in the chat, only the chat as a whole. This might limit

[121] On Community notes, see e.g. : Uku Kangur, Roshni Chakraborty and Rajesh Sharma, "Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes" (arXiv, June 18, 2024) <http://arxiv.org/abs/2406.12444> accessed April 2, 2025; Chiara Patricia Drolsbach, Kirill Solovev and Nicolas Pröllochs, "Community Notes Increase Trust in Fact-Checking on Social Media" (2024) 3 PNAS Nexus pgae217.

[122] In future research, it can be questioned how this applies to TikTok's 'Footnote' affordance once it is rolled out in Europe. See https://newsroom.tiktok.com/footnotes?lang=en

submitting the precise location of the illegal content. By and large however, Twitch' reporting affordances can be considered easily accessible.

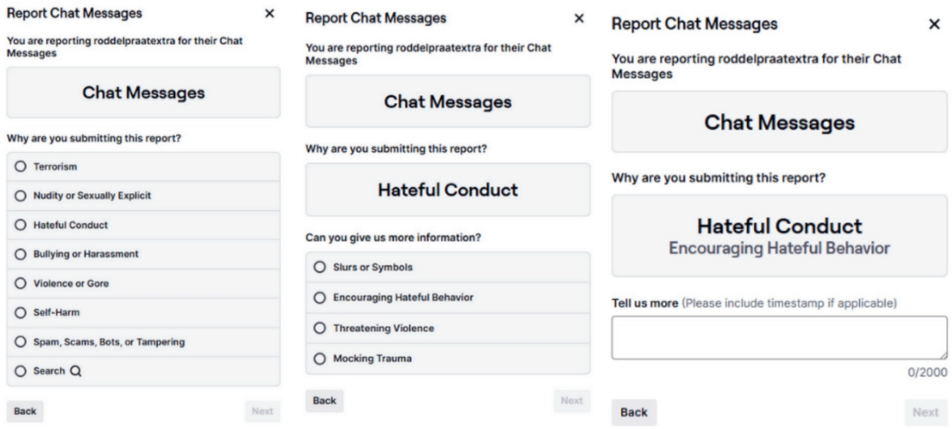


Figure 6: Reporting live chat messages on Twitch

Finally, TikTok offers different access to the reporting mechanism on iOS than it does on its web-interface; on its web-interface the user accesses the notice-and-action mechanism through an ellipsis-menu for posts, comment, livestreaming and livechats, whereas on iOS the user holds down the content to gain access to the notice-and-action takedown mechanism, or uses the 'share' affordance, which is easily findable. Both options are located next to the information that is reported on. However, as noted in table 1, TikTok requires more interactions than the other platforms to reach the article 16 notice-and-action mechanism. This is because their notice-and-action mechanism is integrated as an option in their TaC mechanism; this feature brings us to the next point, namely the relation of the notice-and-action mechanism for terms and conditions to the article 16 mechanism.

The integration of article 16 mechanisms in T&C mechanisms is not precluded by article 16.[123] This is true for Twitch, as it is unclear when content is reported on the basis of its illegality as opposed to its incompliance with terms and conditions (see figure 8). Twitch acknowledges this and states that every notice will be processed as an article 16 notice, but will first be reviewed on the basis of terms and conditions: "upon receiving a report, our moderation team evaluates the reported content against our Community Guidelines and other applicable policies […]. If the content

[123] Wilman, Kalėda and Loewenthal (n 30) 136.

does not violate our Community Guidelines, it is further reviewed for illegality based on the information provided in the report."[124] In its guidance on its reporting flow it also reports that content that does not fall under any of the T&C categories can be reported under an 'Other illegality' category, only available within the European Union.[125] This affordance was not visible throughout the technical walkthrough despite having taken place in the European Union in various member-state language settings, and could only be found using the 'search' function.

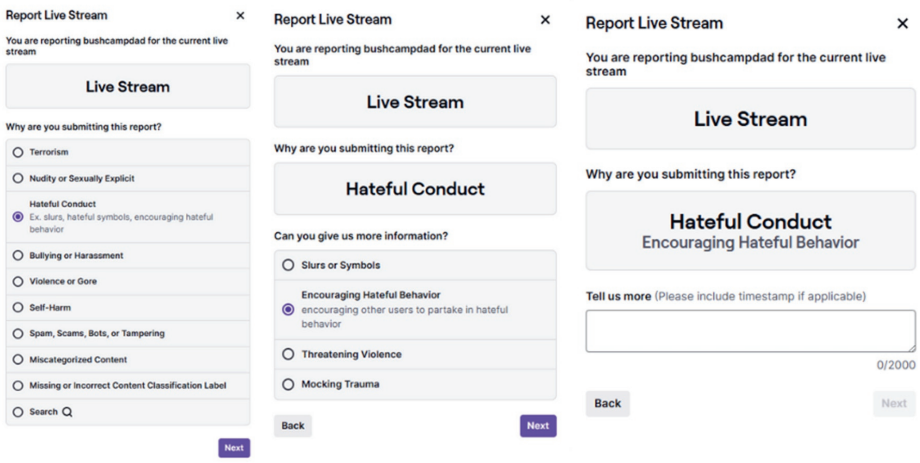


Figure 7: Twitch Notice-and-action flow for reporting hate speech

AliExpress has a similarly integrated reporting mechanism, in which it is unclear whether their reporting mechanisms are aimed at illegal content or TAC (Figure 8). The user can report comments under illegal categories (such as CSAM under articles 3 to 7 Directive 2011/93/EU and criminal offences pursuant to Article 18 DSA), but also more vaguely defined categories such as 'fake'. Products can be reported for IP-infringement and 'Prohibited products', which are both illegal categories. However, their transparency report does not specify under what set of norms complaints fall.

[124] 'Twitch DSA Transparency Report – February 2025', (*Twitch*, 2025) <https://safety.twitch.tv/s/article/Twitch-DSA-Transparency-Report-February-2025?language=en_US> accessed 29 October 2025.

[125] 'Digital Services Act ("DSA"), (*Twitch*, 2025) <https://help.twitch.tv/s/article/digital-services-act?language=en_US> accessed 29 October 2025.

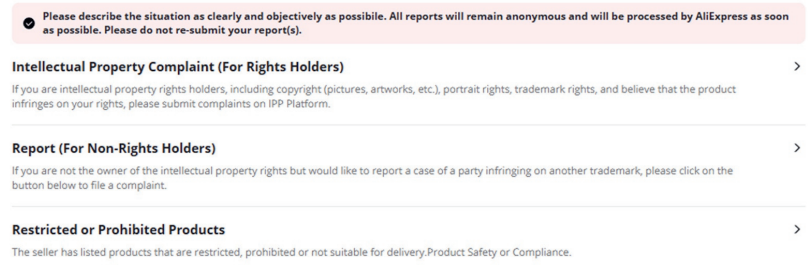


Figure 8: AliExpress reporting options for product

The indistinctness of article 16 mechanisms from TaC-mechanisms is particularly relevant for the question whether the reporting mechanism is equivalently findable as the T&C-mechanisms.[126] Reporting on the basis of the law creates legal certainty, creates consequences for notification, and affects liability, as illustrated in Section 2. Hosting services are therefore expected to place notice-and-action mechanisms equivalently findable to the information as T&C-mechanisms. As noted above, it is difficult to assess this for Twitch and AliExpress, but assuming they have integrated both mechanisms, they are equivalently findable; the question is whether a mechanism is findable if the user is unaware they are using it.

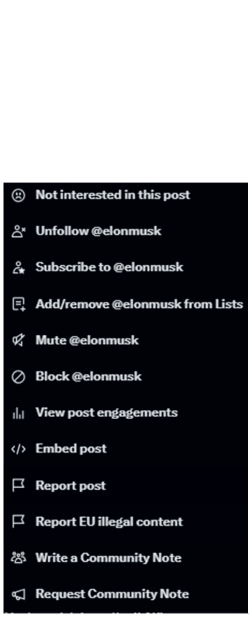


Figure 9 X 'Report EU illegal content' option

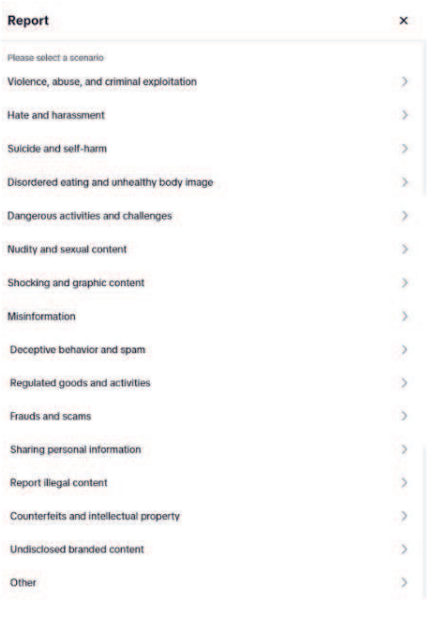


Figure 10: TikTok web-based reporting menu

[126] Regulation 2022/2065, rec. 50.

X has abided by the equivalency requirement very visibly: the 'Report EU illegal content button' is similarly findable as the 'Report post' button leading to the T&C-mechanism on both web-based and iOS interfaces (Figure 9). Although this is equivalently findable, the subsequent notice-and-action mechanism is only usable after the user scrolls past diversions aimed at pointing the user toward other mechanism, such as T&C violations, IP violations, trusted flagger mechanisms, and judicial takedown requests. For TikTok, the Report Illegal Content button is located amongst a wide range of other reporting categories that can divert the user. Their transparency report notes this as an explicitly different mechanism than their T&C mechanism, even though their reporting flow also focuses on reviewing content on the basis of their own policies.[127] To report illegal content on the web interfaces, the user is required to scroll down through a range of options (see Figure 10). Although the iOS interface does not require scrolling, it still requires users to navigate a range of categories that might induce the user to select the T&C-mechanism. When reporting hate speech, for example, the user could be inclined to select 'hate and intimidation' (based on terms and conditions) before scrolling down to find the 'report illegal content' button subsequently offering the 'hate speech' option (based on the law). It is questionable whether this makes the article 16 mechanism equivalently findable as the T&C mechanisms.

4.3 User-friendliness

To be accessible and user-friendly, the user must be able to submit notices through mechanisms that are phrased in accessible language and that are appropriate to the medium and audience that the platform hosts. Section 2.2.2. explained that the language in which the notice-and-action mechanism is offered is ideally the same as the content it is used to report; however, the bare minimum would be to provide a notice-and-action mechanism that is offered in the language that is most spoken by EU-citizens, in this case English. All notice-and-action mechanisms studied are available in English and a variety of member state languages.

As noted above, the accessibility of the language poses the complex balancing act of phrasing mechanisms in language that is simultaneously legally accurate and understandable to the user. Sekwenz et al note that it is difficult to reconcile the two.[128] AliExpress' notice-and-action mechanism, particularly regarding product

[127] 'Digital Services Act: Our fifth transparency report on content moderation in Europe', (*TikTok,* 2025) <https://newsroom.tiktok.com/digital-services-act-our-fifth-transparency-report-on-content-moderation-in-europe?lang=en-150>, accessed 29 October 2025.

[128] Sekwenz, Wagner and Parkin (n 25) 538.

reviews (see Figure 11), is opaque on what exactly is reported. Broad phrasings such as 'fake' or 'illegal items', which overlap as categories, are listed with distinctly legal text ('content reported involves one of the offences referred to in Article 3 to 7 of Directive 2011/93/EU), and might therefore dissuade users from using the reporting mechanism. AliExpress' reporting mechanism for products provides a better explanation for options, and therefore seems more user-friendly (Figure 8). However, it may still feel difficult for non-professional users to navigate, as the subsequent menu for prohibited products does not provide much guidance to the user to frame their notice. (see Figure 12)

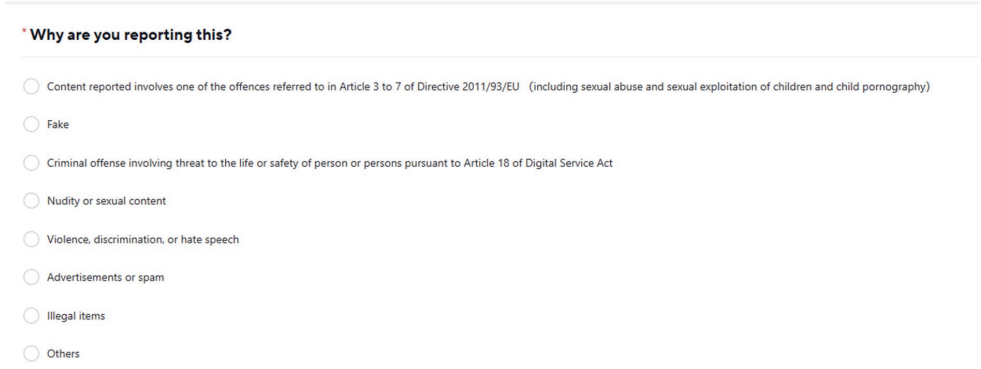


Figure 11 AliExpress reporting options for user reviews

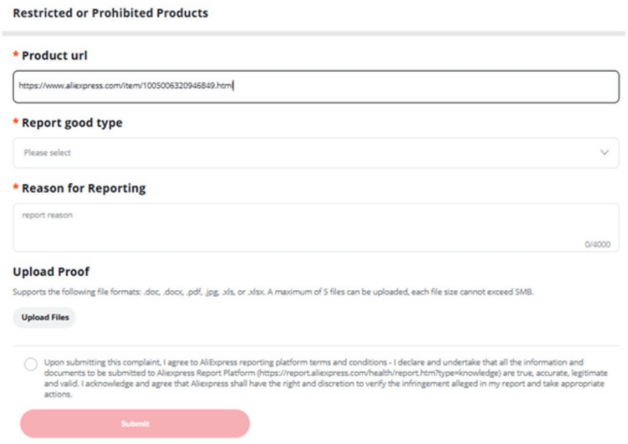


Figure 12 AliExpress reporting prohibited products

X' notice-and-action mechanism uses relatively accessible language. The user receives clear instructions and a relatively comprehensive list of options for illegal content. These are phrased in a relatively open manner.

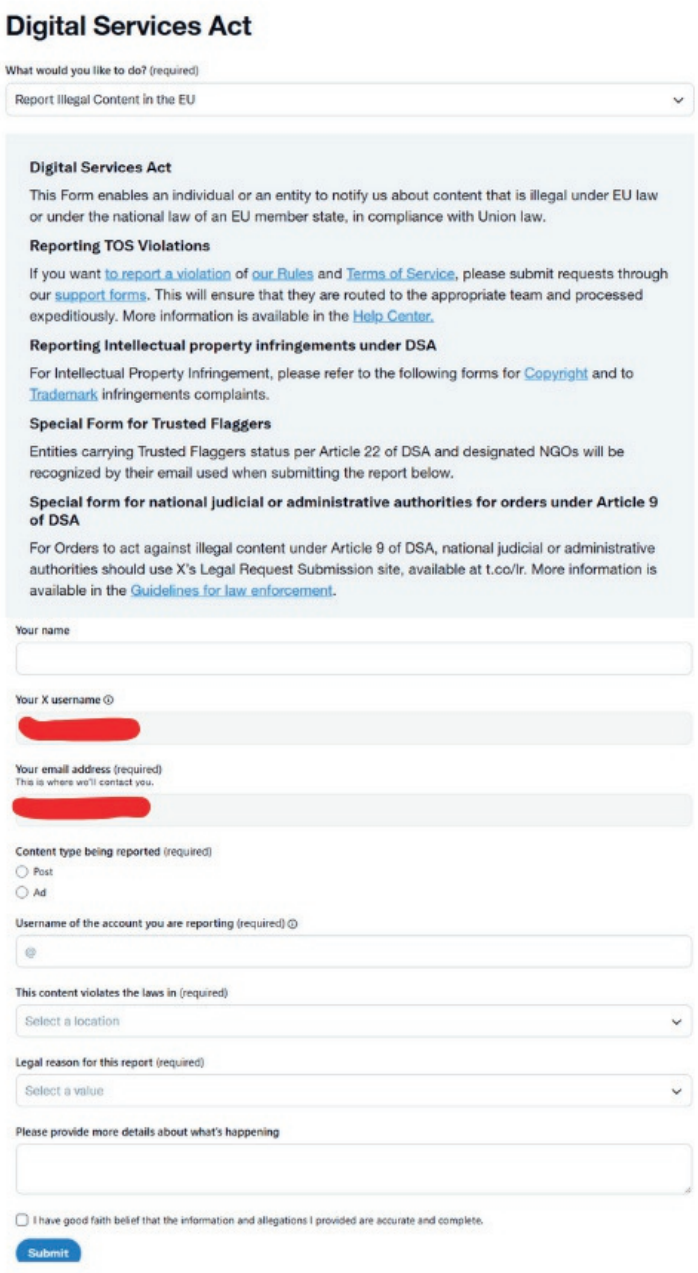


Figure 13 X notice-and-action mechanism

TikTok and Twitch have taken a different approach. Both platforms phrase their notice-and-action mechanisms in relatively accessible manners (see Figures 7 and

10). This is due to the fact that both platforms note in their transparency report that they first review content on the basis of its compliance with terms and conditions, and afterwards on the basis of potential illegality. However, in the case of TikTok, that phrasing may fog the delineation between submitting notices on illegal content and submitting notices on T&C-compliance, which may guide users into a flagging mechanism that does not facilitate the user into providing sufficient information to trigger the liability exemption of article 6. This is exacerbated by the fact that there is significant overlap between the available T&C-categories and illegal content categories. For example, as mentioned above, it is possible to submit a notice on hateful conduct under T&C, and a notice on hate speech under the – less accessible – illegal content menu. Twitch provides the most accessible user language, providing examples of speech that could fall under the listed categories (see Figure 7). However, their language is not clear in indicating whether it is possible to submit notices on illegal content or only based on T&C. The difficulty of user-friendly language in the notice-and-action mechanism illustrates the delicate balance between legal accuracy and catering to a broad user base.

4.4 Facilitates sufficiently precise and adequately substantiated notices

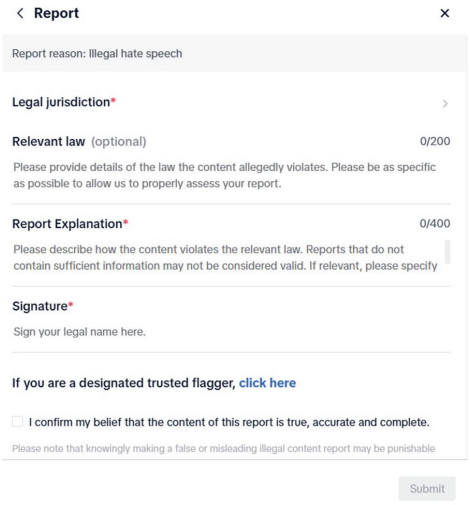
Article 16(2) requires that hosting providers enable the service recipient to submit sufficiently precise and adequately substantiated notices with the aim of triggering the obligation to take down the litigious content or risk liability. This means that the applicant must be able to share a substantiated explanation of reasons, the exact electronic location of the information that is reported, their identity, and a statement of good faith.[129]

On X, the user can submit the jurisdiction they are submitting the notice for, the legal reason for their report, given an unlimited open-text box to explain their issue with the content in light of the legal reason selected, must check a box indicating good faith, and are given the opportunity to share their legal name (Figure 13). The location of the information the user is reporting is prefilled. TikTok offers a similar layout: users can indicate what jurisdiction they seek to report content in, they can identify the relevant law they refer to, explain the violation of the law by that information, *must* share their identity, except for in cases of CSAM (Figure 3), and check a box indicating the notice is submitted in good faith. The requirement for a legal name raises a question in light of recital 50 that suggests that users may submit notices anonymously as explored in Section 2.2.4. Of course, one can submit a fake

[129] Regulation 2022/2065, art. 16(2).

name, which would circumvent the requirement. However, the compulsory nature of the identification requirement is not explicitly required under the DSA.

On the substance of the notice, encompassing the nature of the illegality: X provides a drop-down menu with a list of grounds for illegality, whereas on TikTok the user is invited to refer the relevant law without such categorisation; although this does not nudge the user in any specific direction, it presumes significant legal knowledge on the user's part.[130] TikTok limits the space users have to substantiate their reports to 200 characters to indicate the legal ground they rely on and 400 characters to explain why the content is illegal under that ground. This can create a tension with article 16(2) and (3): it is unlikely that any detailed legal explanation can be given in 400 characters to the extent that it can affect liability exemption under article 6. As indicated above, a detailed legal is not possible in their T&C-mechanism: no information can be provided (see Figures 13 & 14). This is further discussed in Section 5.



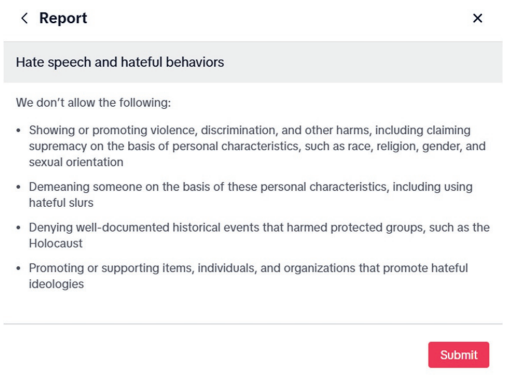


Figure 13: TikTok Article 16: Reporting Illegal Hate Speech *Figure 14: TikTok T&C: Reporting Hate Speech*

AliExpress offers three reporting mechanisms: one for comments, one for products when logged in, and one for products when not logged in. The mechanism for products when logged in allows users to select from a wide range of illegal product categories. Subsequently, the user is given an open-text option with a maximum of 4000 characters to provide information in the 'reason for reporting' (Figure 12). Users

[130] See on this also Sekwenz, Wagner and Parkin (n 25).

may also upload 'evidence' in a variety of document types; for example, in the sphere of product safety, evidence may be required to illustrate why a product is dangerous or defective. Similar to previous platforms, AliExpress requires the user to tick a 'good faith' box.

Twitch offers users less space to provide information. After indicating a category of illegal or harmful content, the user is given one text-box with 2000 characters in which they can indicate how the text was illegal (Figure 7). Whereas the livestream can be flagged with specific timestamps in the notice-and-action mechanisms, indicating the exact electronic location of potentially illegal content in the livechat is difficult; it can only be flagged as a whole. Finally, each of the platforms allows the user to indicate that they submit their notice in good faith. Twitch only allows the user to submit their identity through an email-address, dissimilar to other platforms that allow for full legal name.

4.5 Article 16 compliance

The walkthrough has explored the process of reporting hate speech on AliExpress, TikTok, X, Twitch, and AliExpress. Some interesting trends emerge: the platforms studied in the walkthrough have very different implementations of the notice-and-action mechanism required by article 16, showing that the discretion afforded by the EU regulator is used – and may also be necessary. The following section will outline what this means for compliance, and cross-reference where appropriate to findings from DSA audit reports and the DSA transparency database.

Platform	**Accessible to everyone (art. 16(1))**	**Easily accessible: Location near information that is reported on (art. 16(1))**	**Easily accessible: Location equivalent to T&C mechanism (rec. 52)**	**User-friendly language (art. 16(1))**	**Anonymous Submissions for CSAM (art. 16(2)(c))**	**Sufficiently precise and substantiated notices (art. 16(2))**
AliExpress	Compliant	Limited (in case of product description)	Non-distinct from T&C mechanism	Limited	Compliant (also for other types of content)	Compliant (4000 characters + evidence)

TikTok	Compliant	Compliant	Limited: Added friction compared to T&C mechanism	Compliant	Compliant	Limited (200 characters for legal ground + 400 characters for explanation)
Twitch	Compliant	Compliant	Non-distinct from T&C mechanism	Compliant	Compliant (also for other types of content)	Limited (2000 characters)
X	Compliant	Compliant	Compliant	Compliant	Compliant (also for other types of content)	Compliant (unlimited)

Table 2: Compliance across platforms in three levels: compliant, limited, and non-compliant.[131]

Affordances – unsurprisingly – are tailored to the service offered by the platform. For example, AliExpress focusses its notice-and-action mechanism on product safety and IP, and Twitch emphasising an intuitive notice-and-action mechanism, perhaps at the expense of accuracy due to the speed required to report on illegal content in livestreaming services. This also underscores the point that notice-and-action mechanisms need to be judged in a context-dependent manner. Platforms have uniformly adopted some legal requirements, such as allowing for anonymous submission of reports against CSAM content and allowing users to indicate they are submitting notices in good faith. However, online platforms differ on the implementation of other requirements. Two points stand out: the degree to which the mechanism is easily accessible and equally findable, and the space available for substantiating the notice.

To the first point: the walkthrough shows that the question whether the notice-and-action mechanisms are 'easily accessible' is contentious for AliExpress, TikTok, and Twitch. As indicated above, X' mechanism is clearly identifiable and equally findable as its T&C mechanism.[132] This is confirmed in its audit report. AliExpress' notice-and-action mechanism is located next to the product description, which in principle is close to the product that the notice is aimed at, but the user must scroll

[131] For a more extensive overview please refer to 'Notice and Action features v3' in the repository.

[132] 'DSA Audit Report', (X, 2025) <https://transparency.x.com/en/reports/dsa-audit-report>, accessed 29 October 2025.

down significantly to reach that description and therefore the reporting affordance. There is no indication to the service recipient that the notice-and-action mechanism will be found there, and the product can be purchased without encountering a notice-and-action affordance. This may not be considered easily-accessible. AliExpress' audit report indicates KPMG was unable to assess compliance with article 16(1):[133] the Commission has opened formal investigations regarding this provision, that have ultimately led to binding commitments which have led to the current rendition of the notice-and-action mechanism. The current notice-and-action mechanism is understood to reflect the binding commitments with the European Commission. The question is whether this is compliant with article 16 – if it is, significant scrolling may not necessarily stand in the way of compliance. KPMG also notes that service recipients are unable to submit notices outside of the product level. This may have changed since their audit period as reviews are now explicitly included, but it confirms why reporting hate speech in products specifically can be burdensome in the current notice-and-action mechanism: the mechanism is geared toward product safety, making it difficult to report hate speech in the product pictures, for example.

The ease-of-accessibility can also be a contentious point in case of TikTok. TikTok's notice-and-action mechanism is integrated in their T&C mechanism. In the iOS interface, this requires the user to navigate overlapping T&C categories before reaching the 'report illegal content' affordance further down the list, that may dissuade users from reporting as illegal content. In the web interface, this friction is more pertinent, since it requires the service recipient to scroll down before finding the 'report illegal content' affordance. It is questionable whether this can be considered equivalently findable to the T&C reporting mechanism. However, KPMG's audit report does not note this as incompliant.[134] The added friction can be crucial in the context of platform liability, as the T&C mechanism does not allow for substantiating the notice – and may therefore fail to trigger article 6 DSA. This is further discussed in Section 5.

Twitch' notice-and-action mechanism is easily accessible, but it is unclear whether this is the notice-and-action mechanism for illegal content, as all categories are

[133] 'AliExpress Digital Services Act Audit Report', (*AliExpress*, 2025) <https://www.aliexpress.com/p/transparencycenter/auditReport.html>, accessed 29 October 2025.

[134] 'TikTok DSA Assurance Report', (*TikTok*, 2025) https://sf16-va.tiktokcdn.com/obj/eden-va2/zayvwlY_fjulyhwzuhy[/ljhwZthlaukjlkulzlp/DSA_H2_2024/TikTok-DSA-Audit-Assurance-Report-Sep-2024.pdf , accessed 29 October 2025.

phrased along their terms and conditions. Although Twitch does not have an audit report to refer to, its transparency reporting explains the design choices behind this: all notices are treated as article 16 notices.[135] Once they are submitted, they are first reviewed for compliance with terms and conditions, and subsequently for illegality. This may raise the question whether the user should be aware whether they are reporting content as illegal or as incompliant with terms and conditions. However, since the article 16 mechanism and the T&C mechanism are the same, Twitch' notice-and-action mechanism can be considered easily accessible and equally findable.

The question can be posed what the findability of the notice-and-action mechanism means for the usage of such mechanisms. The transparency reports and transparency database show how many notices are received, how many are framed as article 16 notices, and how many are followed up by the platforms. Table 3 illustrates this. [136]

Platform	Monthly Active Users	Notices Received	Article 16 notices received	Content removed as incompliant	Content removed as illegal
AliExpress	172.1 million	803123	803123	353,268[137]	52787
TikTok	169 million	N/A	308755	15365	26512
Twitch	28.4 million	2654718	2654718	421625	104
X	105 million	N/A	335676	31382	88697

Table 3: Data on notices received and content removed.

Table 3 is noteworthy in light of the analysis presented above for two reasons. Firstly, it shows that Twitch and AliExpress, by treating every notice seemingly as an

[135] 'Twitch DSA Transparency Report – February 2025', (*Twitch*, 2025) https://safety.twitch.tv/s/article/Twitch-DSA-Transparency-Report-February-2025?language=en_US, accessed 29 October 2025.

[136] The transparency reports from AliExpress and X refer to the period October 2024-March 2025. TikTok's transparency report refers from January 2025 to June 2025. Finally, Twitch' transparency report refers from 17 February to 31 December 2024. It can be considered a limitation that the time period do not fully overlap; however, submissions to the DSA transparency database show that it is difficult to discern large variations in trends, see e.g. Gautam Kishore Shahi and others, "A Year of the DSA Transparency Database: What It (Does Not) Reveal About Platform Moderation During the 2024 European Parliament Election" (arXiv, April 9, 2025) <http://arxiv.org/abs/2504.06976> accessed October 10, 2025. In the case of Twitch, the findings need to be contextualised against the longer time period that it is reporting on. It also needs to be noted that various authors have rightfully highlighted the limitations of self-reported numbers; as it stands, there is little enforcement on the accuracy of transparency reporting. See on this: Amaury Trujillo, Tiziano Fagni and Stefano Cresci, "The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media" (2025) 9 Proc. ACM Hum.-Comput. Interact. CSCW187:1; Kaushal and others (n 49); Aleksandra Urman and Mykola Makhortykh, "How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms" (2023) 47 Telecommunications Policy 102477; Mikkel Flyverbom, "Transparency: Mediation and the Management of Visibilities" (2016) 10 International Journal of Communication 110.

[137] Since AliExpress did not specify, this needs to be calculated as 87% (removals for scope of platform service) of 406055 (total number of removals).

article 16 notice, receive the most article 16 notices.[138] This is especially true for Twitch, which has a comparatively smaller number of users. Of course, there can be a number of other reasons for this: there may be more illegal content on AliExpress and Twitch than on X and TikTok, or the reporting user base on those platforms may be larger. Secondly, the data from the transparency reports show that the notification option for illegal content remains important, despite terms and conditions often being the more appealing option for hosting providers; TikTok's data shows that, despite reviewing 308k notices by their terms and conditions, it is still necessary to 'fall back' on illegal content as a ground for removal in 26512 cases as it could not be removed under their terms and conditions. This is especially relevant in light of the consequences that may have for the intermediary liability exemption of article 6 discussed in the next section.

The second point of interest is the ability to substantiate the notice as required by article 16(2)(a). This is also relevant in relation to the separation of T&C and article 16 affordances: when a hosting providers chooses to integrate the T&C and article 16 mechanisms, the service recipient must still be enabled to submit a sufficiently substantiated statement of reasons. For this requirement, X provides unlimited space for service recipients to substantiate their notice. AliExpress limits the space to 4000 characters, but provides service recipients the opportunity to upload evidence. TikTok and Twitch can be considered more limitative: Twitch provides users 2000 characters to explain why content is illegal, TikTok 200 for the legal ground and 400 for explanation. The question is whether this impedes users in providing an adequately substantiated notice: delicate balancing acts around illegal content may not be explained in a 400 characters, but remaining overly broad may not enable the hosting provider to decide to remove content without an independent legal examination. In these cases, the requirements of article 16 create a paradox: whereas TikTok and Twitch arguably have very user-friendly notice-and-action mechanisms, they may be limitative in enabling legal accuracy. While X and AliExpress are considered less-user friendly, they do provide the user with unlimited space for legal accuracy. This highlights the inherent contradiction that the requirements of ease-of-access and user-friendliness pose in relation to the requirements of article 16(2), the liability exemption of article 6(1), and article 16(3). This is further explored in the next section.

[138] This is also true when correcting Twitch' data for the larger timeframe of 10.5 months as opposed to 6 months for the other platforms.

5. Implications of Article 16 Compliance for Platform Liability

This section reflects on how the notice-and-action mechanisms discussed above affect the intermediary liability exemption of article 6 DSA. Liability for hosting illegal content is established through national law or EU law.[139] The DSA harmonises a liability exemption: platforms are not liable for hosting illegal content if they do not have actual knowledge of the illegal content and were not aware of facts or circumstances from which the illegal content is apparent, or upon obtaining such knowledge, act expeditiously to remove or disable access to the illegal content.[140] The barrier for liability for platforms is a relatively steep one, as it balances responsibility for illegal content with potential chilling effect for freedom of expression resulting from over-moderation.[141] To add responsibility without shifting the balance, the DSA creates a system of accountability instead of liability.[142] Notices under article 16 are, as iterated above, the most important avenue for creating liability for user-generated content by triggering the liability exemption of article 6. The notice-and-action mechanisms as laid down in article 16 are a due diligence obligation: it requires platforms to enable users to participate in content moderation. However, although due diligence obligations are generally separate from liability exemption[143], notice-and-action mechanisms are also inseparable from the liability exemption of article 6.

As emphasised in Section 2, the main function within the framework of liability for hosting providers of notice-and-action mechanisms is to create 'knowledge' of illegal content with the platform, thus excluding the platform from liability exemption.[144] Conversely, if a hosting provider is unaware of illegal content, and the notice is inadequately substantiated to the extent that it does not allow the platform to remove the content without an independent assessment, it cannot be held liable.[145] The Court of Justice of the European Union only attributes meaning to imprecise notices to the extent that they can be a factor, amongst others, on the basis of which a national court can decide that a diligent economic operator may have been aware of

[139] Regulation 2022/2065, art. 3(h). See also Wilman (n 29).

[140] Regulation 2022/2065, art. 6(1).

[141] Giancarlo F Frosio, "From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe" (2017) 12 Journal of Intellectual Property Law & Practice 565; Kuczerawy, "Intermediary Liability & Freedom of Expression" (n 21).

[142] See generally Martin Husovec, "Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation Of Global Internet Rules" (2023) 38 Berkeley Technology Law Journal 883.

[143] Regulation 2022/2065, rec. 41.

[144] Regulation 2022/2065, art. 16(3); art. 6(1).

[145] Court of Justice of the European Union, 22 June 2021, *Youtube/Cyando,* Conclusion of A-G Saugmandsgaard Oe, C-682/18 and C-683/18, para 187-188.

illegal content.[146] The quality of notices is therefore crucial.[147] The requirements of article 16(2) for notice-and-action generally reflect the necessary attributes for a sufficiently substantiated notice,[148] although discussion it is debatable whether the identity of the reporter and the statement of bona fide belief are required to trigger an obligation for takedown.[149]

The previous section has identified gaps in compliance with article 16. These gaps create a tension with the liability exemption: (i) when the notice-and-action mechanism of article 16 is located in such a way that it guides users to choose the T&C mechanism, as for example with TikTok, it may guide them into a mechanism that does not allow them to provide sufficiently precise or adequately substantiated notices; (ii) if the notice-and-action mechanism is designed in a way that does not allow the provision of information in line with article 16(2) in a user-friendly manner, it can lead to notices not being sufficiently precise or adequately substantiated. Both scenarios lead to the notices not giving rise to 'actual knowledge' to trigger the liability exemption, and can therefore result in the platform not taking action against this illegal content. Assuming in scenario (i) that the content is addressed under T&C, it may still be removed, avoiding the issue of illegal content remaining available – after all, most content is removed under terms and conditions;[150] however, if content is not removed, it may not be possible to trigger liability, since the platform could argue it was not sufficiently aware of the illegality of the content.

While all platform designs touch in some aspects to this analysis, the most pertinent point affecting liability exemption is visible in AliExpress' and TikTok's mechanism. TikTok places its report illegal content functionality amongst other T&C categories. Users may therefore be inclined to use T&C mechanisms; however, TikTok's T&C mechanism does not provide the same functionality for providing information as the 'report illegal content' options do. Users cannot indicate jurisdiction or provide any explanation on the illegality of the content (compare Figures 14 and 15). Especially in fields where legal assessments are complex, this may require the platform to make

[146] Court of Justice of the European Union, 22 June 2021, *Youtube/Cyando,* C-682/18 and C-683/18, para 115; Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09, para 122.

[147] Wolters and Gellert (n 28) 409–410.

[148] Hofmann and Raue (n 43) 330.

[149] Wilman, Kalėda and Loewenthal (n 30) 139.

[150] Kaushal and others (n 49); Trujillo, Fagni and Cresci (n 137); Daria Dergacheva and others, "One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database" (Center for Media, Communication and Information Research 2023) <https://platform-governance.org/2023/one-day-in-content-moderation-by-social-media-platforms-in-the-eu/> accessed December 14, 2023.

those assessments, which is exactly the scenario A-G Saugmandsgaard Øe sought to avoid in his conclusion in *Youtube/Cyando*.[151] Without being able to provide sufficient information, the user is dependent on whether a national judge takes an imprecise notice into account as a relevant factor to establish whether the platform was aware of illegal content, as established in *L'Oreal v eBay*.[152] In the context of a limited liability regime that requires notices to provide information on the content and the nature of the illegality, this leaves the user in uncertainty.[153] In the context of AliExpress: their notice-and-action mechanism is difficult to find; the user can purchase products without ever encountering one, raising the question whether this is considered easily accessible. A subsequent question is whether these design choices are deliberate to avoid liability by discouraging submitting notices that lead to actual knowledge or awareness of illegal content under article 16(3), and therefore ensuring continued exemption from liability under article 6(1).

For TikTok and Twitch, additional concerns for liability exemption stem from compliance with article 16(2). Their notice-and-action mechanisms are limitative in the amount of information users can submit on the illegality of the platform, only providing textboxes for 400, and 2000 characters respectively to explain jurisdiction, what legal norm is relied upon, and to illustrate the illegality of the content. This can be challenging in complex cases, especially since legal jargon is generally not adjusted to 400 character limits. Although these notice-and-action mechanisms are seemingly aimed at legally empowering users in the fight against illegal content due to their user-friendly nature, they constrain how users can explain the illegality of content and therefore create tension under article 16(2). This creates, as noted above, legal uncertainty on how imprecise notices affect the potential liability of the platform for illegal content. More space may be required to create legal certainty around the use of notice-and-action mechanisms.

6. Conclusion

Mandatory notice-and-action mechanisms in article 16 DSA formalises a long-standing platform practice and emphasises the importance of notice-and-action mechanisms in the European intermediary liability exemption framework. Article 16 formulates requirements for notice-and-action mechanisms that allow for the user-friendly and easily-accessible submission of notices that, when submitted correctly,

[151] Court of Justice of the European Union, 22 June 2021, *Youtube/Cyando,* Conclusion of A-G Saugmandsgaard Oe, C-682/18 and C-683/18, para 189.

[152] Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay,* C-324/09, para 122.

[153] Court of Justice of the European Union, 22 June 2021, *Youtube/Cyando,* C-682/18 and C-683/18, para 115.

force the platform to remove content or risk liability. The walkthrough of the notice-and-action mechanisms of AliExpress, TikTok, Twitch, and X in this article shows that implementation varies across platforms. It further indicates that article 16 can create a paradox in practice, where what is considered user-friendly and easily-accessible under article 16(1) may create a tension with the legal detail that a user is required to submit under article 16(2) in order to create knowledge on illegal content with the hosting provider under article 16(3). Intuitive and easy-to-use mechanisms may not allow the user to provide sufficiently detailed legal information. Conversely, very detailed notice-and-action mechanisms phrased in legal jargon may harm the user-friendliness and accessibility of such mechanisms.

Section 4 illustrates that compliance with the requirements of article 16 varies. AliExpress has a notice-and-action mechanism that seemingly integrates T&C and article 16 notices. Although their notice-and-action mechanism is not easily accessible and is not always phrased in a user-friendly manner, it provides ample space for users to indicate why content is illegal and is therefore compliant with article 16(2). X generally has the most compliant notice-and-action mechanism design across the platforms: it is clearly positioned and separated from T&C mechanisms. However, in spite of its design, it is currently under investigation for how it follows up on those notices in line with requirements under article 16(6) DSA.[154] Twitch' notice-and-action mechanism is integrated with its T&C mechanism. It clarifies that it treats every notice as an article 16 notice, making its mechanism both easily-accessible and user-friendly. The 2000-character limit available for substantiation of notices raises some concerns from the perspective of article 16(2), however. TikTok's notice-and-action mechanism is contentious from the perspective that it is not easily-accessible on the web interface, and may create confusion with its T&C mechanism. It also limits the space users have to indicate why content is illegal, which may frustrate the potential for liability in light of article 6 DSA.

Section 5 has argued that the way notice-and-action mechanisms are designed can affect the degree to which platforms can rely on liability exemption under article 6 DSA. When a platform has knowledge of illegal content it must remove that content or risk liability. Knowledge is dependent on the notice-and-action framework. If that framework is difficult to find, or does not allow users to provide sufficient information for the platform to establish illegality without an independent legal

[154] X has reported no statements of reasons in the DSA transparency database over the course of August and September 2025, indicating that notices are not being treated, or at least content moderation decisions are not being uploaded.

assessment, it is possible that a platform is not deemed aware of the illegal content. This can lead to the hosting provider still being able to rely on the liability exemption of article 6 and being under no obligation to remove the illegal content. Whether article 16 is correctly implemented thus determines whether notice-and-action mechanisms enable users to combat illegal content, or whether the platform steers users into reporting outside of the platform liability framework. To enable users in the fight against illegal content, notice-and-action mechanisms deserve serious enforcement.

Chapter V: Article 22 DSA: Building Trust with Trusted Flaggers

Published as: Jacob van de Kerkhof, 'Article 22 Digital Services Act: Building Trust with Trusted Flaggers' (2025) 14 Internet Policy Review 1.[1]

Abstract: *Trusted flaggers have long played a role in content moderation: a bilateral, voluntary affair between online platforms and individuals or organisations that are afforded prioritised access to the content moderation process. Due to their expertise, they are trusted to 'flag' illegal or harmful content. Article 22 Digital Services Act formalises this framework, allowing governmental and non-governmental organisations to apply for certification as trusted flaggers and requiring online platforms to treat their submitted notices on illegal content with priority and without undue delay. The certification of the first trusted flaggers under Article 22 has sparked public debate about their influence and power, especially in relation to the freedom of expression of internet users. Concerns about trusted flagger frameworks are new in part, but also reflect existing weaknesses in the framework relating to over removal and freedom of expression. This contribution explains the concerns about trusted flagger involvement in content moderation in light of freedom of expression, assesses the promises and pitfalls of Article 22 in its current implementation, and offers recommendations to ensure more effective operationalisation of trusted flaggers under Article 22 and better safeguard the right to freedom of expression of internet users.*

[1] The author is grateful to the participants in the DSA and Platform Regulation Conference 2024 for their feedback on a first presentation of these findings, Dr. Taylor Annabell for her feedback on the final version, and the Internet Policy Review reviewers/editors Valère Ndior, Julian Rossi, Anne Bellon, and Frédéric Dubois for their helpful comments.

Introduction

The selection of the first trusted flaggers under Article 22 Digital Services Act (Regulation 2022/2065, 'DSA') has sparked public debate, fuelled by concerns over their power and potential impact on freedom of expression on the internet.[2] Trusted flaggers are individuals or organisations that are afforded prioritised access to the content moderation process of online platforms. Commonly, they have access to a privileged channel to 'flag' content – submit a notice indicating illegality or incompatibility with community guidelines[3] – which is subsequently addressed with priority by the online platform compared to the treatment of 'ordinary' flags. Access to this privileged flagging mechanism is based on expertise in content areas. Organisations functioning as trusted flaggers range from national enforcement authorities, such as the Dutch Ministry of Internal Affairs,[4] to non-governmental organisations (NGOs), such as the INHOPE network reporting child sexual abuse material (CSAM). Trusted flaggers are commonly seen as positive additions to the content moderation process, but have remained relatively understudied actors.[5] While trusted flaggers act primarily in their own interest, seeking to combat illegal or harmful content,[6] they also support the interests of online platforms, who want to rid their platforms of harmful content to make their space safer and more attractive for users and – crucially – advertisers.[7] However, the role of trusted flaggers in content moderation has also been subject to critique. Namely, that trusted flaggers can be 'censoring' through opaque means, potentially pushing a political narrative and controlling the public discourse by removing unwanted content.[8] This is amplified when a state entity as a trusted flagger submits notices that lead to removal because it implies sidestepping constitutional norms to unduly interfere with freedom of expression.[9] Originally, trusted flagger arrangements were a voluntary effort by online platforms to channel the expertise of civil society into their

[2] E.g. Jan Schneider, 'Woher kommt der Wirbel um "Trusted Flagger"?' (*ZDFheute*, 16 October 2024) <https://www.zdf.de/nachrichten/politik/deutschland/trusted-flagger-bundesnetzagentur-zensurvorwurf-100.html> accessed 4 November 2024.

[3] Kate Crawford and Tarleton Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 New Media & Society 410.

[4] *Kamerstukken* 2022-2023, nr. 1599

[5] Sebastian Felix Schwemer, 'Trusted Notifiers and the Privatization of Online Enforcement' (2019) 35 Computer Law & Security Review 105339.

[6] Naomi Appelman and Paddy Leerssen, 'On Trusted Flaggers' (2022) 24 Yale Journal of Law & Technology 452.

[7] Rachel Griffin, 'From Brand Safety to Suitability: Advertisers in Platform Governance' (2023) 12 Internet Policy Review.

[8] E.g. Winfried Veil, 'Verpestet ist ein ganzes Land … (Teil I)' (*CR-online.de Blog*, 2024) <https://www.cr-online.de/blog/2024/10/16/verpestest-ist-ein-ganzes-land-teil-i/> accessed 4 November 2024.

[9] Jacob van de Kerkhof, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' in José van Dijck and others (eds), *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values* (Amsterdam University Press 2025).

content moderation process. Examples of such programmes include Amazon's Project Zero, Meta's Trusted Flagger programme, and YouTube's Priority Flagger programme. The EU-regulator has adopted this idea in co-regulatory documents such as the Strengthened Code of Practice on Disinformation 2022 (comm. 21) and the Code of Conduct on Countering Illegal Hate Speech Online 2016, encouraging providers of online platforms to offer voluntary trusted flagger frameworks. Similar examples are found on a national level. For example, the now-defunct 2016 German Netzwerkdurchsetzungsgesetz created trusted flagger-like entities called Beschwerdestellen: expert organisations that forward complaints to providers of online social media platforms, such as Jugendschutz.net.

The DSA codifies this formerly voluntary framework in Article 22. The DSA-enforcing authorities at member state-level – Digital Services Coordinators (DSCs) – select trusted flaggers that must be recognised by all providers of online platforms.[10] Once the trusted flagger is awarded their status, providers of online platforms must facilitate notice submissions through the mechanisms created in line with Article 16 DSA, and ensure these are given priority and processed without undue delay. Article 16 DSA requires providers of online platforms to create notice and action mechanisms that allow users to notify them about illegal content, providing an explanation of why the individual thinks the information is illegal, where it can be found, the identity of the user, and a statement of good faith. The contribution of this paper is to (1) situate Article 22 DSA against the background of current critiques of trusted flagger frameworks in light of the protection of freedom of expression in the European Union, (2) identify promises and pitfalls in the new framework offered by Article 22, and (3) propose steps that can be taken to provide better safeguards regarding freedom of expression, a more harmonised implementation of Article 22 and better operationalise the trusted flagger framework. The article is structured as follows: Section I highlights existing critiques of the trusted flagger framework; Section II situates Article 22 against that backdrop and assesses whether it has delivered on promises of legitimacy or raises additional pitfalls; Section III finally proposes concrete measures that can be taken to ensure legitimate and harmonised implementation of Article 22 DSA. It is currently early days regarding the implementation of Article 22: as of January 2025, there are only 16 selected trusted flaggers. This article therefore makes a timely and novel contribution by (a) clarifying the current public debate on trusted flaggers in the DSA context, and (b)

[10] Regulation 2022/2065, art. 22(1).

proposing steps that can be taken by the European Commission to harmonise the application of Article 22 under Article 22(8) – creating the possibility for the Commission to issue further guidance-, improve freedom of expression safeguards, and create a more effective trusted flagger framework.

The article draws on legal and socio-legal sources related to trusted flaggers. The article builds on case law from jurisdictions beyond the EU, by lack of EU case law, to illustrate what tensions may arise within the EU. In terms of the fundamental rights framework drawn upon, this article refers to EU fundamental rights as expressed in the Charter of Fundamental Rights ('CFR'), which can be interpreted in light of case law by the European Court of Human Rights ('ECtHR') following article 52(3) CFR.

Section I: The existing trusted flagger framework and its weaknesses

The existing trusted flagger framework in content moderation is a private initiative, relying on the willingness of online platforms to engage with individuals or organisations with particular expertise to aid in content moderation. In doing so, online platforms provide trusted flaggers with certain privileges in the content moderation process, ranging from dedicated flagging portals and employees to whom complaints can be submitted to holding an advisory role in setting community guidelines. This allows trusted flaggers to represent particular interests that might otherwise be underrepresented, such as those of minorities or vulnerable groups.[11] One example of such groups is the INHOPE network, focussing on combating CSAM content. Although the trusted flagger framework is generally perceived as a positive step toward more effective and legitimate content moderation due to the expertise of the flagger and the potential for civil-society involvement, scholars have also criticised the framework. This section outlines those critiques, against which Article 22 DSA is subsequently evaluated in Section II. The critique centres on three points: (i) risk of over-removal; (ii) fundamental rights and rule of law concerns; and (iii) transparency.

(i) Risk of over-removal

One of the primary concerns in the landscape of content moderation is the risk of over-removal. It has been theorised that internet liability laws would lead platforms to excessively remove content because they would likely 'err on the side of caution'

[11] Appelman and Leerssen (n 6) 467.

in their moderation practices in order to avoid liability.[12] The same argument is levelled at trusted flaggers. A trusted flagger is afforded flagging privileges on the basis of their expertise[13] and the quality of their work.[14] Therefore, their flags might carry more weight, creating a risk for what Schwemer calls 'rubberstamping': a situation in which the online platform restricts content without in-depth review.[15] This is in line with the European Commission's attitude to trusted flaggers, stating that every flag should be treated with "an appropriate degree of confidence as regards their accuracy".[16] Limited platform review becomes problematic when a trusted flagger errs about illegality or incompatibility with terms of service. In such cases, trusted flagger involvement could unduly restrict internet users' speech, impeding their freedom of expression, which is further discussed in the following section. Appelman and Leerssen suggest that this risk for over-removal is driven by the fact that trusted flagger organisations often have specific interests, and they can therefore be aggressive in submitting notices pertaining to content conflicting with those interests.[17] In doing so, their voices can outweigh those that propose a more balanced approach with regard to freedom of expression.[18] Platforms are expected to comply with these flags under the DSA risking over-removal of potentially legal speech.

Empirically, the risk for over-removal is difficult to assess, since online platforms are rarely transparent about the specific content they remove, and information about the proportion of notices submitted that led to removal is even more opaque. Some circumstantial evidence can be found in literature: Eifert and colleagues' evaluation of the NetzDG shows that complaints by complaint bodies (*Beschwerdestellen*) were removed at a marginally higher rate (18% vs 14%) than complaints by individuals.19 This indicates that, although such notices are more likely to be removed, this does not happen to the extent that it should be considered over-removal or censorship. A slight increase in removal rates might be expected because of the expertise of trusted

[12] Brian Chang, 'From Internet Referral Units to International Agreements; Censorship of the Internet by the UK and EU' (2018) 49 Columbia Human Rights Law Review 113, 122–123.

[13] Katharina Kaesling, 'Vertrauen als Topos der Regulierung vertrauenswürdiger Hinweisgeber im Digital Services Act' (2022) 86 UFITA 328, 341.

[14] Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024) 207.

[15] Schwemer (n 5) 8.

[16] Recommendation 2018/334, rec. 29

[17] Appelman and Leerssen (n 6) 468.

[18] Brenda Dvoskin, 'Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance' (2022) 67 Villanova Law Review 447, 469–475.

[19] Martin Eifert (ed), *Netzwerkdurchsetzungsgesetz in Der Bewährung: Juristische Evaluation Und Optimierungspotenzial* (1. Auflage, Nomos 2020) 40.

flaggers in identifying illegal or incompatible content. On the other hand, Urban and colleagues' 2017 study shows that copyright holders submitting notices through privileged avenues has led to a significantly higher rate of removal, and that removal often happened without significant review on the part of the online platform.20 This could indicate over-removal, which would be problematic: the interest of the IP-holder is not the protection of freedom of expression, but a commercially-incentivised protection of copyrighted creations, which might inspire erring on the side of caution in submitting notices, in line with Dvoskin's observation that interests groups advocating more content moderation have outweighed those advocating less. That said, with some exceptions, the illegality of IP-infringing content can be less contentious from a freedom of expression perspective; in cases of piracy or counterfeiting, content flagged can be clearly illegal, whereas in cases of political speech, this can be more complex as hate speech and misinformation are more gray area policy fields.

Some data is also available on the Dutch Ministry of the Interior acting as a trust- ed flagger with Facebook and Twitter.[21] In parliamentary debate, excerpts of takedown requests were shared: the Ministry was flagging disinformation around elections on behalf of municipalities who struggled to gain access to social media platform operators.[22] Misinformation is a legally contentious issue, and the border between legal and illegal misinformation is often unclear. Facebook responded accordingly to this ambiguity, not complying with their flags on disinformation. However, upon government requests about reinstating blocked accounts, Facebook reinstated accounts of government officials and posts thus complying with the Ministry's wishes.[23] While the sample size of these requests is very small (5), it illustrates that in legally contentious areas, platform providers can be more careful with 'rubberstamping', pushing back to government wishes. This suggests that platform review can in potential be an effective safeguard in the trusted flagger framework, and 'rubberstamping' could be overstated. It is difficult to assess whether the trusted flagger framework really leads to over-removal – likely this is dependent on the actor requesting, the nature of the content on which the request is submitted and the incentive for the platform to comply. Even if content is not over-removed, trusted flagger practices can still lead to fundamental right concerns, as explained in the

[20] Jennifer M Urban, Brianna L Schofield and Joe Karaganis, 'Takedown in Two Worlds: An Empirical Analysis' (2017) 64 Journal of the Copyright Society of the USA 483, 504–505.

[21] van de Kerkhof, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' (n 9).

[22] *Kamerstukken* 2022-2023, nr. 914

[23] *Ibid.*

following paragraph.

(ii) Fundamental rights and rule of law concerns

The right to freedom of expression of internet users can be at risk when legal content is removed without justification. In the European Union, freedom of expression is protected under article 11 CFR, which, following article 52(3) CFR, can be interpreted using case law on article 10 of the European Convention on Human Rights (ECHR) of the European Court of Human Rights, since those articles cover the same right. Freedom of expression covers the right to express yourself on the internet,[24] and the right to information.[25] Content being taken down can interfere with that right, either by limiting the speech of the internet user or the right to information of the internet user who can no longer access that content. Interferences to these rights are allowed, if they meet the requirements of article 52(1) CFR, meaning they pursue a legitimate aim, are prescribed by law, and are necessary in a democratic society.[26] Usually, such interferences are subject to judicial review, assessing if the listed requirements are met; when submitting informal requests – 'flags' – such balancing acts are not made by a judicial authority, but by the entity submitting the notice, and ultimately by the platform moderating content. Although the final balancing act is indeed made by the platform, it is possible that pressure from such requests by privileged entities leads to removal.[27] This creates a scenario in which although the platform makes a voluntary decision to remove content, the pressure exuding from a removal request resulting in removal can interfere with the internet user's freedom of expression.

The tension with fundamental rights is present in the case of private entities acting as trusted flaggers as well as government entities acting as trusted flaggers. Private entities are not directly bound by fundamental rights norms,[28] but have a responsibility, implicitly in EU national case law[29] and explicitly in the DSA, e.g. Article 14(4), to respect fundamental rights norms in the provision of their services. For state entities, acting as a trusted flagger can create a more pertinent fundamental

[24] *Melike v Turkey*, 35786/19 (2021), ECtHR.

[25] *Cetin et al v Turkey*, 40153/98 & 40160/98 (2003), ECtHR, para 64.

[26] *The Sunday Times v UK*, 6538/74 (1979), ECtHR, para 49; *Animal Defenders v UK*, 48876/08 (2013), ECtHR, para 100

[27] Chang (n 12) 122–123.

[28] Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (1st edn, Cambridge University Press 2022) 196–200; Claudia E Haupt, 'The Horizontal Effect of Fundamental Rights' in Giovanni De Gregorio, Oreste Pollicino and Peggy Valcke (eds), *Oxford Handbook on Digital Constitutionalism* (Oxford University Press 2024) 10.

[29] *Facebook*, III ZR 179/20, 2021, BGH

right tension.[30] State entities carry particular responsibility with regards to the freedom of expression of internet users; more so than platforms. State entities issuing informal removal requests can interfere with freedom of expression, since their notices likely carry more weight due to the potential regulatory sanctions that can follow from non-compliance.[31]

In various non-EU jurisdictions, courts have deliberated the constitutionality of state entities functioning as trusted flaggers with regards to freedom of expression, illustrating the constitutional strain occurring when governmental authorities engage – often in an opaque, non-accountable manner – in content moderation. For example, the Israeli Supreme Court addressed questions about state entities' privileged access to content moderation in *Adalah Legal Center v State Attorney's Office*. In that case, Adalah had complained about Palestinian freedom of expression being violated due to social media platforms being pressured by flags by the Israeli police cyber department. The Israeli High Court found that such requests do constitute a state act that requires a legal basis, since state actors are repeat players that have coercive regulation available to them.[32] However, as long as it is not shown that a flag by a state agent causes the platform to forgo independent assessment or that the flag impedes a platform's discretion, there is not a direct violation of freedom of expression, the High Court concluded.

In the US, the Supreme Court decided on a similar matter in *Murthy v Missouri*. The case considers dealings between White House officials and social media platforms requesting takedown of Covid-19 misinformation. The 5th Circuit of Appeals had decided that a platform's content moderation practices can be so induced by government conduct that there is coercion or significant encouragement that such conduct can be seen as a First Amendment violation.[33] The Biden Administration had interacted with platforms in a manner that was 'coercive', or at least 'significantly encouraging', affecting the independence of the platform. The Supreme Court did not address the alleged coercion substantively, rather finding that the respondents had no standing to bring suit.[34] However, in his dissent, Justice Alito indicated that he thought that the requests were unconstitutional.[35] Social media

[30] van de Kerkhof, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' (n 9).

[31] Jacob van de Kerkhof, 'Jawboning Content Moderation from a European Perspective' in Charlotte van Oirsouw and others (eds), *European Yearbook of Constitutional Law: Constitutional law in the Digital Era*, vol 5 (1st edn, TMC Asser Press 2024).

[32] Adalah Legal Center for Arab Minority Rights in Israel v State Attorney's Office, Cyber Department, HCJ 7846/19 (Isr.)

[33] Missouri v Biden, 595 U.S. (2022), 5th. Cir., p. 50-51

[34] Murthy v Missouri, 601 U.S. (2024),S. Ct.

[35] Murthy v Missouri, 601 U.S. (2024),S. Ct., Alito dissenting, p. 16

platforms are also acutely aware of the tension arising from government authorities acting as trusted flaggers.

In *UK Drill Music*, the Meta Oversight Board emphasised that law enforcement can provide important context and expertise in the content moderation process, but it is important that Meta platforms make independent assessments based on that information and accurately represents its engagements with law enforcement on these matters.[36] The cases above illustrate that government authorities engaging with content moderation can cause a tension with fundamental rights; however, in the EU, courts are yet to deliberate the freedom of expression concerns associated with the use of trusted flaggers.

Aside from concerns for freedom of expression, in cases where state entities function as trusted flaggers, there can also be concerns for the rule of law, in light of their responsibility to act in accordance with the law. Content moderation is based either on the law or on the basis of platforms' community guidelines. Acting as a trusted flagger provides state entities a privileged channel to submit notices on the basis of community guidelines, creating a rule of law tension: de facto state entities could be enforcing private guidelines, not the law. Under the rule of law, state entities require a legal basis to engage in content moderation, even if it simply means encouraging platforms to enforce their own community guidelines.[37] That legal basis is missing when content is moderated on the basis of its terms of service. It is well-documented that terms of service are the preferred basis of content moderation.[38] Empirically, researchers have confirmed this by studying the DSA Transparency Database holding statements of reasons on content moderation decisions by online platforms.[39] UN Special Rapporteur Kaye provides two reasons why this practice could be problematic when state entities are involved: (i) terms of service are often broadly phrased,[40] meaning that there is little legal certainty of what content is regarded as illegal, and (ii) terms of service agreements often go beyond what is

[36] UK Drill Music, Oversight Board, 2022.

[37] Schwemer (n 5) 10.

[38] Niva Elkin-Koren, Giovanni De Gregorio and Maayan Perer, 'Social Media as Contractual Networks: A Bottom Up Check on Content Moderation' (2022) 107 Iowa Law Review 987; João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, 'How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications' (2023) 48 Computer Law & Security Review 105792.

[39] Rishabh Kaushal and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024); Amaury Trujillo, Tiziano Fagni and Stefano Cresci, 'The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media' (arXiv, 15 December 2023) <http://arxiv.org/abs/2312.10269> accessed 15 January 2024.

[40] See also Danielle Keats Citron, 'Extremist Speech, Compelled Conformity, and Censorship Creep' (2018) 93 Notre Dame Law Review 1035, 1058.

prohibited by law.[41] State entities using terms of service to flag content can de facto sidestep legal requirements to request content removal. Eghbariah and Metwally flag this issue, citing instances of police internet referral units – police departments that focus on investigating malicious content on the internet – wrongly targeting extremist content online relying on broad definitions in terms of services resulting in unwarranted censorship.[42] This creates a paradox: states and platforms rely on definitional ambiguity in terms of service and lack of scrutiny present in trusted flagger frameworks because they are a low-cost, highly expedient way of targeting not only unlawful, but also legal but harmful content on the internet, which may ultimately benefit society as a whole, but comes at the expense of strict adherence to fundamental rights safeguards.[43]

(iii) Transparency

The paragraphs above have alluded to the current trusted flagger framework being opaque. (Lack of) transparency is an oft-discussed topic in content moderation. The general consensus is that content moderation is an opaque process, increasing the accountability deficit in regulating online platforms (Suzor, 2019; Leerssen, 2020, 2024; Gill, 2021; Rieder & Hofmann, 2020).[44] Increasing transparency about content moderation practices is suggested to potentially increase accountability in the content moderation process,[45] but the operationalisation of transparency has inspired critique on allowing online platforms to control the discourse over their practices,[46]

---

[41] David Kaye, 'Report of the Special Rapporteur on the Promotion of the Right to Freedom of Opinion and Expression' (Human Rights Council 2016) A/HRC/32/38 14–15 <https://primarysources.brillonline.com/browse/human-rights-documents-online/promotion-and-protection-of-all-human-rights-civil-political-economic-social-and-cultural-rights-including-the-right-to-development;hrdhrd99702016149> accessed 11 November 2024.

[42] Rabea Eghbariah and Amre Metwally, 'Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service' (2021) 23 Yale Journal of Law & Technology 542, 587–589.

[43] Schwemer (n 5) 12.

[44] Nicolas Suzor and others, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 1526; Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (2020) 11 European Journal of Law and Technology 51; Paddy Leerssen, 'Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act' (2024) 4 Weizenbaum Journal of the Digital Society; Karanjot Gill, 'Regulating Platform's Invisible Hand: Content Moderation Policies and Processes' (2021) 21 Wake Forest Journal of Business and Intellectual Property Law 171; Bernhard Rieder and Jeanette Hofmann, 'Towards Platform Observability' (2020) 9 Internet Policy Review.

[45] Hannah Bloch-Wehba, 'Global Platform Governance: Private Power in the Shadow of the State' (2019) 72 SMU Law Review 28; Rachel Griffin, 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability' (2023) 14 Transnational Legal Theory 46; David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (1st edn, Columbia Global Reports 2019).

[46] Mikkel Flyverbom, 'Transparency: Mediation and the Management of Visibilities' (2016) 10 International Journal of Communication 110.

yielding inconsistent results,[47] and suffering from a general lack of applicability to the process of algorithmic content moderation due to the opaque nature of automated moderation.[48]

Private arrangements of platforms with entities operating as trusted flaggers are similarly opaque. Platforms do not need to be forward about whom they afford flagging privileges, leaving users unable to control what actors are involved in content moderation. Platforms, meanwhile, control the discourse about their trusted flagger framework by being forward about their interactions with NGOs, but not about their interactions with governmental organisations or IP-right organisations, which are generally less popular.[49] In practice, this is confirmed in the Oversight Board decision *UK Drill Music*. The Oversight Board found that transparency reports underrepresent interactions with law enforcement, raising concern since trusted flaggers can hold a flagging bias to which feedback is not available. This relates to submissions made by trusted flaggers, but also to trusted flagger practices as a whole. The funding, organisational structure, and (in)dependency of trusted flaggers cannot be assessed if it is unknown who these organisations are, leaving little room for accountability. This lack of transparency creates an accountability deficit, which can harm the legitimacy of the trusted flagger framework. Schwemer, building on Suchman's approach to legitimacy as 'a generalised perception of desirability',[50] theorises that legitimacy is also "about being representative" of a group, the general public, or a specific interest.[51] In the current opaque scenario, there is no way of holding organisations accountable for the desirability of their actions, practices, and the interests they represent, meaning that this detracts from the legitimacy of the trusted flagger framework.

Section II: Article 22 DSA

This section positions the framework of Article 22 against the backdrop of current

[47] Eleni Kosta and Magda Brewczyńska, 'Government Access to User Data: Towards More Meaningful Transparency Reports' in Rosa-Maria Ballardini and Olli Pitkänen (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer Law International 2019); Aleksandra Urman and Mykola Makhortykh, 'How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms' (2023) 47 Telecommunications Policy 102477.

[48] Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973; Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18.

[49] Appelman and Leerssen (n 6) 470–471.

[50] Mark C Suchman, 'Managing Legitimacy: Strategic and Institutional Approaches' (1995) 20 The Academy of Management Review 571.

[51] Schwemer (n 5) 8.

critiques of trusted flagger frameworks, identifying the promises and pitfalls of the DSA trusted flagger system. The DSA is a newly entered-into-force EU Regulation that aims – in part – to ensure better protection of fundamental rights online. Article 22 DSA establishes that providers of online platforms must take measures to ensure that trusted flaggers, as appointed by the DSCs, can submit notices through the mechanisms referred to in Article 16 DSA, and those notices are processed with priority and without undue delay. Article 22 departs from private trusted flagger arrangements in which online platforms bilaterally engage with trusted flaggers, and forms a logical next step after a series of nudges in the European Union that stimulate online platforms to include trusted flaggers in the content moderation process.[52] The framework creates a formal avenue for trusted flaggers to gain privileged access to content moderation with all providers of online platforms active in the EU. This can benefit smaller organisations that have struggled to develop a relationship and gain access to a platform, but also help larger organisations gain access to platforms that currently do not have a trusted flagger framework, for example X.[53] The compulsory nature of the Article 22 trusted flagger framework is particularly relevant in a landscape where companies such as Alphabet and Meta are increasingly withdrawing from content moderation efforts in search of political favour.[54] Regulatory pressure ensures the continued participation of selected trusted flaggers in the content moderation process.

The DSC appoints trusted flaggers based on their (i) particular expertise (ii) independence from any platform provider and (iii) diligence, accuracy and objectivity.[55] Trusted flaggers can be governmental and non-governmental organisations, but not individuals.[56] DSCs must avoid appointing too many trusted flaggers, which could decrease the value of the framework, hence precluding individual rights holders from applying.[57] Trusted flaggers must publish yearly reports with the DSC of the notices they have submitted with platform providers. In the event that a trusted flagger submits a significant number of imprecise, inaccurate or unsubstantiated notices, the DSC may, after being informed by the provider of an

[52] Husovec (n 14) 207; Schwemer (n 5) 10–11.

[53] Sheila Dang, 'Twitter Dissolves Trust and Safety Council' *Reuters* (13 December 2022) <https://www.reuters.com/technology/twitter-dissolves-trust-safety-council-2022-12-13/> accessed 18 July 2025.

[54] Jacob van de Kerkhof, 'Musk, Techbrocracy, and Free Speech' (*Verfassungsblog*, 17 January 2025) <https://verfassungsblog.de/musk-techbrocracy-and-free-speech/> accessed 24 March 2025.

[55] Regulation 2022/2065, art. 22(2).

[56] Regulation 2022/2065, rec. 61.

[57] Benjamin Raue, 'Artikel 22 Vertrauenswürdige Hinweigeber' in Franz Hofmann and others (eds), *Digital Services Act: Gesetz über digitale Dienste* (1. Auflage, Nomos 2023) 398.

online platform, open investigations on that trusted flagger and simultaneously suspend the organisation. Investigation can lead to revocation of trusted flagger status. It is important to note that Article 22 is without prejudice to existing trusted flagger arrangements:[58] existing arrangements can continue to exist. This can create a tension that is addressed in subsection II.ii.

(i) Risk of over-removal

Article 22 DSA forms both a solution and exacerbator to the risk of over-removal as a consequence of the trusted flagger system, as discussed in the previous section. Because Article 22 grants certified trusted flaggers access to all online platforms, trusted flaggers with sufficient resources have a wider range of operations than they previously had – potentially enlarging the risk for over-removal. However, the DSA mitigates the risk of over-removal in the trusted flagger framework at four instances: the trusted flagger, online platforms, DSC, and user.

Firstly, the trusted flagger has to report on its activities yearly under Article 22(3), including the number of notices accompanied by the type of illegal content and the action taken by the provider. There is accountability in transparency: transparency reports will show if a trusted flagger overextends its competence by submitting notices that do not address illegal content in the sense of article 3(h), and thus do not require action by online platforms. Over-removal is mitigated further by the requirement for accuracy and expertise in the certification of trusted flaggers: DSCs have to select organisations that have demonstrable expertise in their field, and therefore certified organisations are less likely to submit incorrect notices.

Secondly, online platforms also have a defence against trusted flaggers over-submitting notices: Article 23(2) DSA gives online platforms the right to suspend flagging privileges if entities frequently submit notices that are manifestly unfounded – this includes trusted flaggers.[59] Similarly, if notices are structurally unfounded because they are overturned in internal appeal procedures under Article 20(4) DSA online platforms can request the DSC to suspend a trusted flagger under Article 22(6) DSA. Thirdly, the DSC can revoke the status of a trusted flagger if it is over-submitting under Article 22(7) DSA. Fourthly, the user of the online platform is provided with a range of possibilities for appeal if they find that their content is wrongly moderated in Article 20 and Article 21 DSA. This, in a general sense,

[58] Regulation 2022/2065, rec. 62.

[59] Raue (n 57) 405.

mitigates the risk for over-removal: in the previous bilateral trusted flagger structure, the user had no means to revert a content moderation decision – the DSA provides those means, thus somewhat countervailing the power of the trusted flagger.

Section I.i proposes that the risk of over-removal arises from trusted flaggers representing specific interests, leading to aggressive flagging practices that do not necessarily balance freedom of expression with the interest they represent. This risk can be present in the interpretation of the selection criteria for trusted flaggers under Article 22 DSA. The selection criteria of Article 22(2) DSA are multi-interpretable, and can be coloured based on national preference, since trusted flaggers are ultimately certified by national DSCs. In an EU landscape where rule of law backsliding exists[60] and attitudes toward speech vary regionally, it is possible that organisations that are selected could form a risk to freedom of expression.[61] Such organisations could potentially over-submit notices with online platforms, leading to over-removal. The selection criteria as they are laid down in article 22 do not provide an effective backstop to this: trusted flaggers are selected on particular expertise (Art. 22(2)(a)), implying they will have a particular interest in combatting a specific type of illegal content. Article 22(2)(c) then specifies that the trusted flagger must be able to carry out its activities diligently, accurately and objectively. 'Objectively' as a requirement could form a safeguard to ensure that the trusted flagger weighs their interests with that of the freedom of expression of internet users. However, that criterion is, in the selection procedure, generally demonstrated through submitting a range of successful notices as evidence.[62] Unsuccessful notices would generally be flagged by the platform, who, as explained above, has little interest in disagreeing with the trusted flagger's notices. It is therefore difficult to gauge whether the requirement of 'objectivity' ex Article 22(2)(c) provides a sufficient counterbalance to the explicit interests that trusted flaggers have that could result in over-removal.

Next to the multi-interpretable norms for selection, the DSA does not specify the means trusted flaggers must use to detect content. While some trusted flaggers rely on hotlines or their own investigation representing a more artisanal, case-by-case

[60] Kim Lane Scheppele, 'The Treaties Without a Guardian: The European Commission and the Rule of Law' (2023) 29 Columbia Journal of European Law 93.

[61] Jenny Orlando-Salling and Louisa Bartolo, 'The Digital Services Act as Seen from the European Periphery' (*DSA Observatory*, 5 October 2023) <https://dsa-observatory.eu/2023/10/05/the-digital-services-act-as-seen-from-the-european-periphery/> accessed 18 November 2024.

[62] Raue (n 57) 399.

approach to content moderation,[63] others may look to use algorithmic means to detect illegal content: this is common practice for entities detecting CSAM or copyright infringing content. The risks for freedom of expression related to algorithmic content moderation have been well documented.[64] The use of algorithmic content moderation for detection by trusted flaggers raises the same concerns: is it desirable that the scalability of their activities is achieved at the expense of accuracy? Using automated means also raises a transparency concern: online platforms are required to indicate whether they use automated means to detect illegal content and whether they rely on automated decision-making in the statement of reasons they send to users under Article 17 DSA, but if the flag is submitted by a trusted flagger the user is not informed of any automated detection mechanisms used, de facto sidestepping the requirement of Article 17 DSA.

Aside from remedies pertaining to specific removal of content, the user has no means to challenge the privileged status of a trusted flagger. Trusted flaggers are selected by the DSC, a process which is subject to national administrative law. Whether it is possible for interested parties to appeal such a decision depends on the administrative law of the member state, fragmenting the degree to which users are able to have redress against the selection of trusted flaggers. This, coupled with the fact that trusted flagger statuses can only be revoked by the DSC that awarded the status, makes it challenging for users who seek to challenge a trusted flagger established in a different member state.

(ii) Fundamental rights and rule of law concerns

Section I.ii outlines the fundamental rights risks of the existing trusted flagger framework. Those risks predominantly pertain to state entities flagging as trusted flaggers, as their interference with freedom of expression without providing sufficient justification under article 52(1) CFR is problematic from a fundamental rights standpoint. The concern for fundamental rights is accompanied by rule of law concerns related to flagging on the basis of terms of service. Article 22 DSA reduces both risks, but there are some caveats.

---

[63] Robyn Caplan, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute 2018).

[64] Federica Casarosa, 'When the Algorithm Is Not Fully Reliable: The Collaboration between Technology and Humans in the Fight against Hate Speech' in Hans-W Micklitz and others (eds), *Constitutional Challenges in the Algorithmic Society* (1st edn, Cambridge University Press 2021); Celine Castets-Renardt, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) 2020 University of Illinois Journal of Law Technology & Policy 283; Robert Gorwa, Reuben Binns and Christian Katzenbach, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945.

First of all, the number of state entities seeking certification under Article 22 is tempered somewhat by Article 9 DSA, which provides a clear avenue for administrative and judicial authorities seeking to order the takedown of illegal content – they can do so directly with the platform with more legal safeguards in place – as well as the sentiment expressed in DSA recital 61 suggesting that the number of trusted flaggers should be kept to a minimum. The primary concern in Section I.ii was that state entities acting as trusted flaggers can submit notices that carry enough weight for platforms to remove content without proper review. By doing so, the state entity is able to indirectly remove content without providing a legal basis, a legitimate aim, and weighing the necessity in a democratic society – which are the requirements for interfering with the freedom of expression. Under Article 22, this risk is mitigated: trusted flaggers can only flag content through the mechanisms of Article 16 DSA. Article 16 DSA requires that notices must be submitted on the basis of illegal content, which is defined in article 3(h) DSA as "any information that […] is not in compliance with Union law or the law of any Member State which is in compliance with Union law […]." This provides a degree of legal certainty to users, and decreases the concern of state entities sidestepping the requirements of 52(1) CFR and article 10(2) ECHR: certified trusted flaggers can only submit notices on content that is illegal – and not when it is non-compliant with terms of service.

Despite the additional layer of accountability, there are three loopholes to the operationalisation of flagging through Article 16 that form a risk to the rule of law, not only for state entities but for all trusted flaggers. Firstly, although Article 16 specifies that notices can only be submitted for illegal content, appointed trusted flaggers could still de facto flag on the basis of terms of service, just not within the framework of Article 16. If a notice is submitted outside of Article 16, there is no need for transparency reporting on such notices under Article 22(3). Conversely, it is not expected that online platforms provide the same priority and expedience to these notices. However, Recital 62 hints that online platforms should provide similar (priority) treatment to notices submitted by entities that are not awarded trusted flagger status to "take quick and reliable action against content that is incompatible with their terms and conditions". It is unclear how this sentiment relates to the requirement for priority under Article 22 – after all, priority can only be given once[65] – but it shows that trusted flagger mechanisms can *de facto* be used to flag on the

[65] Raue (n 57).

basis of terms and conditions. Only *de jure*, it cannot under Article 22 DSA.

Secondly, the expected implementation by online platforms of Article 22 raises concerns: some platforms that have a framework will simply onboard newly certified trusted flaggers into the existing framework. Amazon, Facebook and Instagram have pledged that their companies will include new trusted flaggers within their existing ecosystems,[66] which allows for flagging under terms of services under the standards of Article 14 DSA. This is somewhat in the spirit of Recital 62, suggesting that trusted flaggers outside of Article 22 DSA can be treated similarly as those certified by a DSC. However, such portals usually allow for flagging both on the basis of terms of service as well as the law, maintaining the risk for the rule of law.

A third loophole is found in the backstop of Article 20(4) DSA, which serves as a requirement for DSCs to start investigations on trusted flaggers per Article 22(6) DSA. Article 20 describes the DSA's internal complaint handling system: when a user finds that their content is wrongly restricted, they can submit a complaint through the online platform's complaint handling system. If the platform finds that the content is not illegal and not incompatible with the platform's terms and conditions, it reverses its content moderation decision. However, it may be that the flagged content is not illegal but is incompatible with terms and conditions. For example, consider the case when a trusted flagger has submitted a notice per Article 16 on content it considers hate speech, which the platform acts upon and restricts the user's content. When the user submits a complaint through Article 20, and the platform reassesses its content moderation decision, it finds the content was not illegal under national hate speech law, but is still incompatible with the platform's terms and conditions. The content in that case is still removed, but the legal basis for its removal has changed – it is no longer removed on the basis of the law, but solely on the basis of terms of service, thus creating tension with the rule of law safeguard anchored in illegal content in Article 16.

Another point under the DSA trusted flagger framework that forms a challenge to the rule of law is the transnational nature of the internet. Content that is illegal in one member state can be legal in another, but trusted flaggers combatting illegal content receive their status for the entire EU. This can lead to the removal of content

[66] 'Conference's notes - Getting Ready for the DSA: The Role of Trusted Flaggers for a Better Moderation of Platforms' (*Régulation du Numérique*, 8 February 2024) <https://regulation-tech.cnam.fr/summary-arcom-event-on-january-26th-2024-getting-ready-for-the-dsa-the-role-of-trusted-flaggers-for-a-better-moderation-of-platforms/> accessed 14 March 2024.

that is potentially legal in a given member state but illegal in another.[67] For example: LGBT+ related content in advertisements is illegal in Hungary,[68] but not in other EU member states.[69] A Hungarian trusted flagger could flag this content outside of Hungary, leading to its removal by lack of a vigilant online platform, even though the legal basis in that member state is missing. The exercise of determining whether the legal basis for the notice submitted is valid in another member state is left to the discretion of the trusted flagger and the discretion of the platform, who ultimately decides whether the content is restricted.[70] Both are not necessarily equipped to check whether the legal basis for removal contravenes the sentiment of Article 3(h) DSA that national law must be in compliance with EU law before content is actually illegal. Therefore, the legal certainty of users can be at risk when trusted flaggers can de facto start enforcing their national law across other member states. Platforms would be encouraged to use geo-blocking as an adequate remedy, ensuring that content remains available where it is legal.[71]

(iii) Transparency

The previous section has explained that the current trusted flagger framework suffers from an opacity problem. As hinted above, Article 22 DSA provides improvements in that regard. A primary benefit lies in knowing what organisations have trusted flagger access to content moderation processes: the European Commission maintains a database with accredited trusted flaggers.[72] This allows the identification of trusted flagger organisations, which adds to their legitimacy as part of Schwemer's representation perspective. Additionally, trusted flaggers need to publish reports with the DSC on notices they submit to online platforms in a yearly report. That report shows the number of notices, the nature of the content reported, and the action taken by the platform following the notice.[73] This enables scrutiny over trusted flaggers' activities: do they flag content within their mandate based on their expertise? Are they over/under submitting?

Scrutiny can ultimately lead to suspension or revoking of trusted flagger status under Article 22(6) and (7). Transparency reporting also enables scrutiny over the

[67] Franziska Katharina Mauritz, 'To Define Is Just to Define' (*Verfassungsblog*, 8 April 2024) <https://verfassungsblog.de/to-define-is-just-to-define/> accessed 20 August 2024.

[68] Act LXXIX of 2021

[69] *European Commission v Hungary*, C-769/22 (2022), CJEU

[70] 'Roundtable on the Digital Services Act' (Central European Digital Media Observatory 2024).

[71] Mark Lemley, 'The Splinternet' (2021) 70 Duke Law Journal 1397.

[72] Regulation 2022/2065, art. 22(5).

[73] Regulation 2022/2065, art. 22(3).

platform's response to trusted flagger activities: the remedy attached to the notice, as well as the response time, which indicate how well platforms comply with Article 22 DSA. The possible scrutiny of trusted flagger practices also adds to the legitimacy of the trusted flagger framework, since, in line with Suchman's notion of legitimacy, this allows the assessment of the desirability of their actions. The DSA's intention is for trusted flaggers to positively influence content moderation by ensuring that they help platforms take action against illegal content more quickly and reliably.[74] That intention is dependent on the goodwill of platforms to engage with those trusted flaggers; if online platforms do not respond to trusted flaggers, content will remain unchallenged. Transparency reporting by trusted flaggers on how platforms respond to the notices they have submitted allows DSCs to scrutinise platforms and potentially start an investigation or infringement procedures, thus forming a crucial redress instrument for trusted flaggers against uncooperative online platforms.

However, the DSA has not required providers of online platforms to share in their statement of reasons provided to users whether their restricted content was flagged by a trusted flagger.[75] Therefore, direct accountability of a trusted flagger by the general public is difficult: they will not know whether their content is flagged, and, therefore, are unable to hold the trusted flagger to account. This makes it impossible to privately enforce the DSA-rights and even GDPR-rights (a point beyond the scope of this article, but trusted flaggers can qualify as data processors) against trusted flaggers. This also creates a caveat in the liability for wrongful moderation by online platforms. In principle, one can hold platforms liable for wrongfully moderating content, either through out-of-court dispute settlement bodies ex Article 21 DSA, or through a national court. Online platforms have plenty of incentive to abide by requests of trusted flaggers: they can rely on the expertise of specialised entities which benefit from the legitimacy of being selected by national authorities, a deficit of the trusted flagger framework before Article 22. Conversely, platforms also face potential sanctions in cases of non-compliance. But, what happens if a user challenges the wrongful moderation[76] of content based on the requests of a trusted flagger? Will the online platform remain fully liable as the actor that ultimately removes content? Or is that liability somewhat tempered by the involvement of a trusted flagger? Without users being able to identify trusted flaggers as the source of notices against their content, they are unable to place the liability – at least partly –

[74] Regulation 2022/2065, rec. 61.

[75] Regulation 2022/2065, art. 17(3)(b).

[76] Michael Klos, 'Wrongful Moderation' [2020] Nederlands Juristenblad.

with that trusted flagger, if necessary. Of course, there can be good reasons to maintain the anonymity of the flagging entity: flagging content for removal can be experienced as an antagonistic action[77] which may lead to adverse actions against the trusted flagging organisations. However, since trusted flaggers are publicly certified and their identity is accessible via the Commission's database, it makes sense to include their identity in statements of reasons, in order to allow scrutiny of a publicly certified organisation.

Section III: How can we better safeguard freedom of expression in – and ensure effective operationalisation of article 22 DSA?

Despite the implementation of Article 22 being in its infancy, already some recommendations can be made to how it is operationalised and to ensure adequate safeguards for freedom of expression. An initial suggestion was made in the previous paragraph: include the identity of the trusted flagger on the statement of reasons submitted to the user under Article 17 DSA when their content is restricted, to enhance transparency and allow scrutiny of trusted flagger practices. This section proposes additional possible steps that address some deficiencies in Article 22 in the context of Article 22(8) DSA, in which the EU regulator is empowered to further specify how Article 22(2), (6) and (7) must be applied. This article offers three specific recommendations: the first relates to guidance on the application of Article 22(2), the second relates to funding and independence, the third relates to the operationalisation of Article 16 for certified trusted flaggers under Article 22.

(i) Harmonising selection criteria and procedure

Firstly, on selecting trusted flaggers and the requirements for certification. The Board of DSCs – a board made up of all DSCs and the European Commission established in Article 61 DSA – could create standard procedures on the appointment of trusted flaggers, particularly on the timeframes and appealability of such decisions, to equally facilitate organisations across the EU. The application procedure could be harmonised as well. Currently, it is fragmented: in some jurisdictions, for example Finland, organisations seeking certification can get into direct contact with the DSC to submit their application, whereas in other jurisdictions, such as Denmark and Austria, they must use a form. A standardised form that ensures that similar documentation is required from organisations across

[77] Sarah Myers West, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 New Media & Society 4366, 4373.

the EU would be a logical next step. Accompanying the development of such a form, the Board should release guidelines on the interpretation of requirements such as 'independence' and 'expertise' to apply a similar standard of proof across the EU. Independence as a requirement pertains to whether trusted flaggers can be influenced by online platforms in their operation – such would forego their neutral position in content moderation. However, in practice, some trusted flaggers are funded partly by online platforms. The best-known example is the INHOPE network, which is partly funded by big tech companies such as Meta and TikTok. The degree to which such funding is allowed under Article 22(2) is contentious: funding by a tech company could be read as a contraindication of independence. However, safeguarded by a strong organisational structure, operational independence can still be guaranteed. For example, the Austrian DSC has certified two entities that are partly funded by online platforms, because the funding was not deemed significant enough to harm independence.[78] To facilitate trusted flaggers in combatting illegal content, DSCs must take into account funding opportunities for their work, and consider that funding does not necessarily interfere with the principle of independence. The Board should therefore provide guidance on how organisational structures for trusted flaggers can ensure independence while allowing for trusted flaggers to be sufficiently funded, also by industry partners. Finally, the Board should introduce a 'backstop' in the granting of trusted flagger statuses, in cases when selected organisations could form a challenge to the rule of law, for example, organisations that are expected to flag content on the basis of national law that is not compliant with EU law. An example is a trusted flagger certified in Hungary to flag content contravening Hungarian anti-LGBT legislation: illegal content in that member state would not be illegal in the rest of the EU. This 'backstop' could also ensure safeguarding of the 'objectivity requirement' of Article 22(2)(c) discussed above. Using the backstop, the Board has a tool to ensure that trusted flaggers are not overly exercising their privilege at the expense of a balanced approach to freedom of expression. This could be by either limiting the number of trusted flaggers altogether, or by (re)assigning trusted flagger statuses to entities that represent an underrepresented interest, for example in the unlikely scenario that only IP organisations would be selected as trusted flaggers. Such standards can be better safeguarded if the Board has a backstop to revoke trusted flagger status, for example, by request of a majority of the DSCs, in lieu of the accrediting DSC being solely responsible.

[78] KommAustria KOA 16.400/24-017; KOA 16.400/24-013.

(ii) Funding

A second improvement that could be made to the Article 22 DSA framework is funding. As indicated under III.i, funding can be a contentious issue due to the independence requirements trusted flaggers have under Article 22. However, in order to operate, trusted flaggers require resources. A prime candidate to supply funding should be the providers of online platforms – ultimately beneficiaries of their services since trusted flaggers contribute to the detection of illegal content. Although providers of online platforms fund some trusted flaggers already, it is worthwhile investigating whether certified trusted flaggers could benefit from a general support fund managed by the European Commission. A model for this is already present in the DSA: Very Large Online Platforms and Very Large Online Search Engines provide a supervisory fee to fund the Commission's oversight role under Article 43 DSA. In addition to providing better resources to trusted flaggers to combat illegal content, restructuring funding can form an incentive for providers of online platforms to better engage with trusted flaggers, since they are the ones funding their operation. To operationalise this structure, the Commission could set up a fund requiring VLOPs to contribute and divide it equally across certified trusted flaggers. This also creates an incentive for entities to apply for certification with their DSC. Funding restructuring could therefore improve the power balance between the trusted flagger and the platform.

(iii) Creating a harmonised API for trusted flaggers

Thirdly, the DSA leaves platforms discretion to choose the mechanisms for trusted flaggers in line with Article 16 DSA. For example, the Dutch DSC suggests in their DSA guidance that online platforms can facilitate trusted flaggers' mission by appointing staff to deal with notices, or create a designated web interface. The latter approach reflects how some platforms have suggested they comply with Article.[79] In principle, requirements for submitting notices as a trusted flagger are harmonised under the requirements of Article 16 DSA. However, in practice the mechanisms created to comply with Article 16 can differ significantly according to the platform.[80] Earlier research in the context of the NetzDG has shown that the design of flagging mechanisms can affect moderation practices.[81] A standardised application

[79] 'Conference's notes - Getting Ready for the DSA: The Role of Trusted Flaggers for a Better Moderation of Platforms' (n 66).

[80] E.g. Daniel Holznagel, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (2024) 25 Computer Law Review International 172.

[81] Amilie Pia Heldt, 'Bypassing Classical Concepts of Censorship?' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 56, 12.

programming interface ('API') across platforms for trusted flaggers could mitigate some of the fundamental rights concerns raised above relating to the privilege with which those entities flag. It has been suggested that standardising APIs could also be used to ensure DSA compliance.[82] In creating an API for flagging, it is possible to standardise requirements for flagging content across platforms beyond Article 16 DSA[83] whilst simultaneously ensuring compliance of those flags with requirements necessary for limiting freedom of expression. A standardised API can further be used for transparency reporting, since data that needs to be reported on can simply be exported from the API. In cases of state entities acting as trusted flaggers, the API can require additional input that resembles the requirements needed for interference with freedom of expression: legitimate aim, prescribed by law and necessity in a democratic society.[84] Given that legitimate aims are exhaustively listed under art. 10(2) ECHR,[85] it should be possible for state entities to select such aims through this portal. Prescription by law can also be a required part of the notice, both for state and non-state entities. This ensures both compliance with the requirement to only flag illegal content under Article 16 DSA and the requirement of basis in law under Article 10(2) ECHR, also allowing users to assess on the basis of what national law their content is being restricted.

The intended remedy the trusted flagger seeks can also be a requirement in this standardised API. Interferences with freedom of expression must be necessary in a democratic society, which is a requirement of proportionality.[86] Such proportionality can only be assessed if the sought remedy is indicated. Content moderation remedies are diverse, and go far beyond simple 'keep up' or 'take down'. Demonetisation, demoting, and labelling are all effective measures to target unwanted content.[87] In some instances, it can be more effective to label content than to take it down. In contentious cases, where the illegality is not a certainty, this can ensure that freedom of expression is respected while simultaneously ensuring that harmful content is combatted.[88] The proportionality assessment of the trusted flagger should also be

[82] Catalina Goanta, Thales Bertaglia and Adriana Iamnitchi, 'The Case for a Legal Compliance API for the Enforcement of the EU's Digital Services Act on Social Media Platforms' 9.

[83] Husovec (n 14) 461.

[84] Art. 52(1) CFR & 10(2) ECHR

[85] Janneke Gerards, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023) 100.

[86] *Animal Defenders International v the United Kingdom*, 48876/08, (2013), ECtHR, para 100

[87] Eric Goldman, 'Content Moderation Remedies' (2021) 28 Michigan Technology Law Review 1.

[88] Garrett Morrow and others, 'The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation' (2022) 73 Journal of the Association for Information Science and Technology 1365.

shared with the user whose content is affected, ensuring they are informed and able to seek redress.

The harmonisation of flagging portals and inclusion of fundamental rights standards can be difficult to realise from a practical perspective, but the DSA provides a legal basis for the Commission and the Board to promote the development of a harmonised flagging API.[89] It would require providers of online platforms to agree to be part of the same infrastructure. Such would interfere with their freedom to run a business, and require significant resources. From the perspective of the trusted flagger, elaborate requirements in flagging APIs can also be difficult and inhibit expedience. Trusted flaggers have submitted 300.000 statements of reasons over the past 30 days (per 19 November 2024).[90] Combating illegal content online is a matter of expediency and scalability. It has been proposed that this is why the direct application of fundamental rights norms to content moderation is not practical – fundamental rights assessments do not scale to the enormous volume content moderation requires.[91] A trusted flagger portal requiring extensive documentation would, therefore, only intensify the scalability critiques raised in the context of trusted flaggers. To remedy this, it is possible to limit the extensive documentation requirements in the harmonised API to government entities certified as trusted flaggers under Article 22, as that is where the most significant risk for fundamental rights lies. Dutch oversight authorities have reported submitting up to 500 notices per year,[92] suggesting that elaborate documentation, at least for government entities, is achievable.

Conclusion

This contribution has identified existing and newly arisen risks of the trusted flagger framework, and proposed measures to optimise trusted flagger involvement in content moderation under Article 22 DSA whilst introducing further safeguards around freedom of expression of internet users. Relying on trusted flaggers is a proven method for online platforms to help legitimise content moderation practices. Previously, trusted flagger frameworks were bilateral partnerships between

[89] Regulation 2022/2065, art. 44(1)(c).

[90] The DSA transparency database does not allow for making a distinction as to whether those include all trusted flaggers or only Article 22 certified trusted flaggers.

[91] Evelyn Douek, 'The Limits of International Law in Content Moderation' (2020) 6 U.C. Irvine Journal of International, Transnational and Comparative Law 37; Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939.

[92] *Kamerstukken* 2022-2023, nr. 1599.

governmental- and non-governmental organisations and platforms. Such arrangements, while ultimately regarded as beneficial to the content moderation process, suffer from transparency, rule of law and freedom of expression concerns – the latter two particularly in the case of involvement of state entities. Article 22 DSA formalises the trusted flagger framework, creating certified trusted flaggers for the EU. The framework of Article 22 leads to more transparency and legitimacy of trusted flaggers, and helps trusted flagger organisations in the operationalisation of their practices. However, Article 22 has not fully addressed concerns for over-removal, freedom of expression, and rule of law. Additionally, different applications of Article 22 across member states can lead to a diffuse trusted flagger landscape in the EU, exacerbating freedom of expression concerns. This article proposes ways in which Article 22(8) can be used to harmonise the application of Article 22 throughout the EU. Further, the European Commission should investigate whether it is possible to create a uniform means for flagging content for trusted flaggers. In order to fully channel the potential of trusted flaggers, the Commission could set up a funding structure funded by VLOPs to ensure that trusted flaggers can contribute more effectively while simultaneously creating a compliance incentive. Using trusted flaggers is a proven way to expeditiously target illegal content based on the expertise of civil society and government entities; this expediency of removing content may not come at the cost of disregarding safeguards around potential interferences with the right to freedom of expression, especially when practical and easily attainable solutions are available.

Part III: The (Supra)State

"My approach will always be to regulate as little as possible, as much as needed. But if we remain an open continent, it needs to be on our terms, with us in the driving seat."

Thierry Breton, 6 September 2022

Chapter VI: The DSA's Tower of Babel: on the Digital Services Coordinator and Freedom of Expression

Published as: Jacob van de Kerkhof, 'The DSA's Tower of Babel: On Digital Services Coordinators and Freedom of Expression' [2025] European Journal of Risk Regulation 1-26.[1]

Abstract: *The Digital Services Act (Regulation 2022/2065, 'DSA') creates a new national administrative authority to enforce the DSA across member states: the Digital Services Coordinator ('DSC'). DSCs perform a linchpin role in the DSA enforcement. DSCs have a number of tasks that interact with the content moderation process, such as certifying trusted flaggers or participating in drafting codes of conduct. They also have significant investigatory- and sanctioning powers to enforce the DSA vis-à-vis platforms, shaping content moderation processes and protect users' rights against platform misconduct. These interactions with content moderation affect users' freedom of expression. This contribution scrutinises the role of the DSC in light of that freedom, describing how DSCs shape freedom of expression online through their powers in the DSA, and identifying instances where exercise of DSA powers can lead to different levels of protection for freedom of expression across Member States in the decentralised enforcement network. Finally, it proposes avenues in the DSA to anchor protection of freedom of expression in the application of the DSA by DSCs, through pursuing centralisation in cases with significant fundamental rights impact, and encouraging better usage of guideline competencies.*

[1] The author is grateful to Catalina Goanta, Peter Blok, Janneke Gerards, Viktorija Morozovaite, Lisette Mustert, and Nadya Purtova, and participants of the RESHUFFLE Young Scholars Workshop for comments on earlier drafts and insightful conversations on this topic. All mistakes are my own.

I. Introduction

In an effort to combat harmful content online and shift the power imbalance in speech governance, the European Union adopted Regulation 2022/2065, the Digital Services Act ('DSA').[2] The DSA marks a watershed moment in platform governance, adding a significant accountability and transparency framework onto the existing liability framework of the e-Commerce Directive.[3] The DSA promises to ensure a safe, predictable and trusted online environment in which fundamental rights enshrined in the Charter of Fundamental Rights in the European Union ('CFR') are better protected.[4] The DSA creates a layered approach to platform obligations, increasing duties as the size of the platform increases. Platforms with the most reach, and therefore the most potential for harms, have the most obligations under the DSA.[5] Very Large Online platforms and Very Large Online Search Engines ('VLOPSEs', platforms with a number of monthly active users exceeding 45 million)[6] need to abide by a range of risk-mitigating and transparency-building measures,[7] whereas micro- and small enterprises are excluded from the scope of a number of DSA provisions.[8]

In addition to creating a multi-layered regime for platform obligations, the DSA introduces a new administrative authority at the national level, tasked with overseeing DSA compliance: the Digital Services Coordinator ('DSC').[9] The DSC has a two-dimensional relationship with the right to freedom of expression of European internet users in enforcing the DSA. The first dimension is that DSCs can shape the freedom of expression of internet users through drafting norms for content moderation practices and empowering civil society actors to influence the content moderation process along the lines of national laws and speech traditions; although there is a common core of freedom of expression in Europe, there are regional

[2] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), *OJ L 277 27.10.2022,* p. 1-102.

[3] 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,* p. 1-16. See on the DSA in general: Martin Husovec, *Principles of the Digital Services Act* (Oxford University Press 2024); Folkert Wilman, 'The Digital Services Act (DSA): An Overview'; Giovanni De Gregorio and Pietro Dunn, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473, 483–488.

[4] Charter of Fundamental Rights of the European Union, *OJ C 326, 26.10.2012,* p. 391-407.

[5] Regulation 2022/2065, rec. 76.

[6] Regulation 2022/2065, art. 33(1).

[7] Regulation 2022/2065, Section 5, arts. 33-43.

[8] As defined in Recommendation 2003/361/EC; Regulation 2022/2065, art. 15(2); art. 19(1), art. 29(1).

[9] Regulation 2022/2065, art. 49(1)&(2).

differences. The second dimension is that DSCs protect the right to freedom of expression of internet users by acting as a complaint platform for internet users, and by sanctioning platforms for non-compliance. Their role is therefore also to rebalance the power in the online sphere between internet users, providers of intermediary services, and state actors.[10] If enforcement protects the freedom of expression of users, differences in enforcement can lead to different levels of protection of that freedom for internet users in the EU.

This article explores the role of the DSC in the DSA in light of its two-dimensional impact on the freedom of expression of internet users in the EU. It asserts that, while freedom of expression of internet users is anchored to an extent in DSA enforcement, there is a seeming disconnect between the impact DSCs can have on freedom of expression and the safeguards around their position. To this end, it is structured as follows: the first section describes the position of the DSC in the DSA and how its powers interact with the freedom of expression of internet users. The second section argues that the decentralised enforcement structure of the DSA affects the protection of freedom of expression of internet users. The third section makes suggestions as to how freedom of expression can be better safeguarded in the DSCs role in enforcing the DSA.

II. The Digital Services Coordinator

This section dissects the role of the DSC in the DSA, and outlines how its powers interact with the freedom of expression of internet users. Subsection 1 describes the role, competence, and institutional embedding of DSCs, Subsection 2 provides a brief introduction into the legal framework for freedom of expression, Subsection 3 explores how the DSCs' powers interact with freedom of expression, and finally Subsection 4 provides an interim conclusion.

1. DSCs as an institution and their competence

DSCs are national administrative authorities tasked with enforcing the DSA in a member state.[11] Member States may choose to create a new enforcement authority to fulfil the role of DSC, but can also appoint an existing authority, or even divide the enforcement tasks over multiple existing authorities under one 'coordinating' DSC.[12]

[10] On the balancing of power, see generally: Giovanni De Gregorio, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (1st edn, Cambridge University Press 2022); Giulia Gentile, 'Between Online and Offline Due Process: The Digital Services Act' in Annegret Engel, Xavier Groussot and Gunnar Thor Petursson (eds), *New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights* (Springer Nature Switzerland 2025).

[11] Regulation 2022/2065, art. 49(1)&(2).

[12] Regulation 2022/2065, art. 49(2), rec. 110.

DSCs – as decentral[13] enforcers of EU law – are embedded in the procedures of the DSA, but also in national administrative law, which lays down the procedural framework around its supervisory tasks and powers. This contribution touches upon that framework, but predominantly from the perspective of the DSA.[14] Currently, Member States have appointed existing consumer-, media-, and telecom-protection authorities as DSC. DSCs must perform their DSA tasks in an impartial and timely manner.[15] Member States shall provide them with sufficient technical, financial, and human resources to fulfil their oversight- and enforcement tasks.[16] DSCs must fulfil their tasks with complete independence, without prejudice to judicial review or budgetary control.[17] The principle of independence in this regard is contentious. Independence of DSCs could be necessary since they concern themselves with freedom of expression on the internet;[18] independence limits the procedural autonomy of Member States, however, and in the case of the DSA is not derived from primary law.[19] Independence in the DSA is therefore 'rather thin'.[20] The possibility of appointing different, existing, actors, the oversight role of the Commission (which is inherently not an independent authority[21]), and the various mandatory cooperation mechanisms all limit the supposed required of the DSC. An

[13] Generally, on centralised and decentralised enforcement: Robert Schütze, 'From Rome to Lisbon: "Executive Federalism" in the (New) European Union' (2010) 47 Common Market Law Review 1385.

[14] Assessing each DSC in the context of its own jurisdiction would be beyond the scope of this article. In addition to that, not all member states have fully adopted the national implementation laws attributing powers to the DSC, such as Spain and Bulgaria, or have even formally appointed a DSC, such as Belgium and Poland, see on this: https://digital-strategy.ec.eruopea.eu/en/policies/ds-dscs.

[15] Regulation 2022/2065, art. 50(1).

[16] *Ibid.*

[17] Regulation 2022/2065, art. 50(2). See also Pieter Van Cleynenbreugel, 'Independent Administrative Authorities and the Sectoral "Europeanisation" of National Administrative Law: A European Union Law Perspective' (2023) <https://orbi.uliege.be/bitstream/2268/313933/1/EUDAIMONIA_WP20241.pdf> accessed 17 October 2024. The standard that needs to be applied in this case is unclear. On defining 'independence', there is a high standard in the context of data protection Regulation 2016/679, art. 52; Court of Justice of the European Union, judgment of 16 October 2012, *European Commission v Republic of Austria*, C-614/10; Court of Justice of the European Union, judgment of 9 March 2010, *European Commission v Federal Republic of Germany*, C-518/07. In the context of competition law, there is a different standard, see for example Court of Justice of the European Union, judgment of 13 June 2018, *Commission v Poland*, C-530/16. It appears that the Commission had intended for DSCs to not necessarily be independent from other authorities, *see* Regulation 2022/2065, rec. 110, also Gerald Spindler, 'Der Referentenentwurf für ein Digitales Dienste Gesetz (DDG-RefE) – good bye TMG und NetzDG: Erster Überblick über die wesentlichen Weichenstellungen' (2023) 39 Computer und Recht 602, para 35, referring to Matthias Cornlis, Lena Auler and Samuel Kirsch, 'Vollzug des Digital Services Act in Deutschland – Implementierung einer verbraucherorientierten Aufsichtsbehördenstruktur' (Mainzer Medieninstitut 2022) 41.

[18] Timo Rademacher in Franz Hofmann and Benjamin Raue (eds), *Digital Services Act: Article-by-Article Commentary* (First edition, Nomos 2025) 852–857.

[19] As is the case for data protection for example, see art. 16(2) Treaty on the Functioning of the European Union.

[20] Rademacher in Hofmann and Raue (n 18) 853.

[21] The Greens had rejected the proposal to put the Commission in charge of VLOP supervision as it is a political organ. See Husovec (n 3) 425 fn 39.

independent EU agency for oversight was proposed, but eventually rejected because "it stood in the way of expediency of the legislative process;"[22] i.e. harmonisation on aspects of enforcement was not desired at this stage.

DSCs are responsible for the oversight of providers of intermediary services with their main establishment in that Member State.[23] DSCs have competence when an intermediary service provider located in their jurisdiction violates the DSA, when an inhabitant of that member state complains about non-compliance of providers of intermediary services under the DSA, and when it receives complaints from another jurisdiction about an intermediary service provider located in its jurisdiction. DSCs rely on cooperation with other DSCs when enforcing against providers of intermediary services outside of their jurisdiction.[24] Competence as regulated in the DSA is without prejudice to Union law and national rules on private international law, which can aid in determining which DSC is relevant in a particular case.[25] However, the cross-border nature of the internet and the issues addressed in the DSA require an EU-wide approach.[26] The European Board of Digital Services is formed to streamline that cooperation.[27] The Board consists of high-level officials from each Member State DSC, and is chaired by the European Commission.[28] It contributes to a consistent application of the DSA across the European Union – a feature discussed more extensively in Section III.

Decentralised enforcement is a deliberate choice: it ensures that procedures take place at close proximity to citizens of member states, and avoids unnecessary extension of the Commission's competence. That said, the EU regulator codified an exception to this: VLOPSEs. VLOPSEs are challenging to oversee due to their size and limited enforcement power at a national level.[29] Further, the EU regulator may have feared regulatory capture if VLOPSE enforcement was located in a single member state, potentially endangering the required harmonisation and cooperation

[22] ibid.

[23] Regulation 2022/2065, art. 56(1). 'Main establishment' is defined in the DSA as "where the provider has its head office or registered office within which the principal financial functions and operational control are exercised" (rec. 123.) It mirrors the definition of 'place of central administration' *ex* article 4(16)(a) of Regulation 2016/679. See also Opinion 04/2024 on the notion of main establishment of a controller in the European Union under Article 4(16)(a) GDPR, 13 February 2024.

[24] See for example Regulation 2022/2065, arts. 57, 58.

[25] Regulation 2022/2065, rec. 126.

[26] Regulation 2022/2065, rec. 130-131.

[27] Regulation 2022/2065, art. 61(2)(a).

[28] Regulation 2022/2065, art. 62(1)(2); the Commission does not have voting rights (61(3)).

[29] Regulation 2022/2065, rec. 124-125; the enforcement power of the Commission is inflated in this regard due to its supervisory body being paid for by designated VLOPSEs through article 43; national supervisory bodies remain state-funded.

necessary to ensure a safe online environment.[30] After a VLOPSE is designated as such, DSCs share competence with the Commission.[31] The Commission mitigates potential large-scale harms within the EU arising from VLOPSEs by (partly) enforcing the DSA for intermediary service providers with more than 45 million monthly active users.[32] The Commission has exclusive competence over obligations laid down in Chapter III Section 5, containing predominantly transparency-and risk-mitigation obligations.[33] The Commission may also enforce other obligations under the DSA against VLOPs and VLOSEs when it deems appropriate;[34] this could be in the case of systemic infringement of the DSA that seriously affects recipients in a Member State.[35] Once the Commission has initiated proceedings under article 65(1), Member States can no longer start proceedings on the same infringement, to avoid *ne bis in idem*.[36]

At first glance, this reduces the impact of DSCs on freedom of expression; after all, most content is disseminated through VLOPSEs, for which competence is partially vertically centralised with the Commission. However, the DSC Database maintained by European Digital Rights shows that nearly all DSCs have received complaints under article 53 DSA for enforcement against online platforms, and the Irish DSC has requested information from 12 intermediary service providers.[37] In cases of complaints, article 53 does not provide the possibility submitting a complaint with the Commission,[38] only to the DSC of establishment; therefore, for all cases outside of Chapter III Section V, DSCs are the enforcing actor toward VLOPSEs for individual user complaints.[39] The fact that for all obligations outside of Chapter III Section 5 the Commission's enforcement role is limited procedurally, due to the required appropriateness for its involvement, as well as substantively, being that provider oversight in many cases requires significant knowledge of the impact of services on society and significant "political-cultural know-how",[40] supports the role of the DSC

[30] Regulation 2022/2065, rec. 155.

[31] Regulation 2022/2065, art. 56(3).

[32] Rademacher in Hofmann and Raue (n 18) 991.

[33] Regulation 2022/2065, art. 56(2).

[34] Regulation 2022/2065, art. 56(4), rec. 125.

[35] Regulation 2022/2065, art. 65(2).

[36] Regulation 2022/2065, art. 56(4), rec. 125.

[37] 'The Digital Services Coordinators Database' (*DSC Database*) <https://dscdb.edri.org/> accessed 5 May 2025.

[38] Regulation 2022/2065, art. 53: "[…] the Digital Services Coordinator shall assess the complaint and, where appropriate, transmit it to the *Digital Services Coordinator* of establishment […]" [emphasis added]. *Cf.* for obligations under Chapter III Section V see art. 65(2).

[39] Rademacher in Hofmann and Raue (n 18) 944.

[40] Rademacher in ibid 991.

as a key actor in DSA enforcement. Recent cases, such as the Belgian investigation into Telegram,[41] a messaging service noteworthy for its spreading of illegal content that does not yet qualify as a VLOP, or Twitch, a streaming service located in Luxembourg associated with livestreaming public shootings, show that enforcement on the national level by DSCs is highly relevant for platforms just below the VLOPSE threshold.[42] Conversely, inquiries fuelled by the Romanian DSC into the role of TikTok in Romanian elections illustrate the relevance of DSCs as a signalling entity, even when it comes to enforcement against VLOPSEs: the Romanian elections were annulled because of alleged interference through TikTok.[43] This resulted in the Romanian DSC requesting the blocking of TikTok in Romania,[44] and the European Commission launching an official investigation.[45] Individual cases and the Commission's significant, but limited competence, underscore the need for freedom of expression safeguards around the role of DSCs as national enforcers of the DSA.

2. Freedom of Expression

EU citizens enjoy a similar right to freedom of expression under article 11 CFR[46] on the internet as they do in the 'offline world', with the notable distinction that, because of the nature of the internet, harms can be exacerbated by increased reach and longer availability.[47] Correspondingly, the right to freedom of expression requires speakers to take into account the wide audience they reach by communicating online: freedom of expression on the internet is a complex balancing

[41] 'Telegram Still Doesn't Meet Large Platform Requirements under DSA' <https://sg.news.yahoo.com/telegram-still-doesnt-meet-large-095311777.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGA_0t6lA009T3PlmMHDv1K228kVtDkXBlkuappqM2CuEPAlK-YLwk0wp5XpT96_vaubJGnEV1UPgbwFNEBQsyAQ7o2LyJE8V8Y5na55OlvFvRvSUQEKMWwxHUZzHcvzrMzAVkyJforkxdgsxHWdAI6LoHmo7upC6Icn_8fnfQMi> accessed 1 November 2024.

[42] 'Twitch Says Livestream of Buffalo Mass Shooting Was Removed in Less than 2 Minutes | CNN Business' <https://edition.cnn.com/2022/05/15/business/twitch-livestream-buffalo-massacre/> accessed 29 April 2025.

[43] 'Romanian Authorities Ask EU Commission to Investigate TikTok Role in Election | Euronews' <https://www.euronews.com/next/2024/11/27/romanian-authorities-ask-eu-commission-to-investigate-tiktok-role-in-election> accessed 3 December 2024.

[44] Jon Henley, 'Romania Regulator Calls for TikTok Suspension amid Vote Interference Fears' *The Guardian* (27 November 2024) <https://www.theguardian.com/world/2024/nov/27/romanian-regulator-tiktok-suspended-cyber-interference-election-georgescu> accessed 14 January 2025.

[45] 'Commission opens formal proceedings against TikTok under DSA' (*European Commission - European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487> accessed 18 December 2024.

[46] The right to freedom of expression in this contribution is interpreted in light of both Article 11 of the Charter of Fundamental Rights of the European Union, which is used as a reference point in Regulation 2022/2065, and in light of the European Convention of Human Rights; case law by the European Court of Human Rights can be used as a tool for interpretation following article 52(3) Charter of Fundamental Rights of the European Union.

[47] European Court of Human Rights, 18 May 2021, *OOO Informatsionnoye Agentstvo Tambov-Inform v Russia*, 43351/12, para 84; European Court of Human Rights, 25 November 2021, *Biancardi v Italy*, 77419/15, para 62.

act between the rights of the speaker on the internet and the societal interests concerned with being exposed to that speech on the internet.[48] Aside from an active right to freedom of expression, article 11 CFR also encompasses a 'passive' right to receive information.[49] This also encompasses right to access the internet.[50] As such, interferences with access to (parts of) the internet must be balanced with the freedom to access information.[51]

The scope of freedom of expression on the internet is determined predominantly through content moderation. Content moderation is the process by which providers of intermediary services determine what can be expressed and seen on the internet through removing content, amplification and de-amplification, and design of affordances.[52] The process involves a network of stakeholders affecting content moderation, such as civil society, users, and government officials.[53] This process is primarily governed by the intermediary service provider, creating a complicated landscape in which private actors determine the scope of public values such as freedom of expression.[54] This creates a tension: fundamental rights norms, such as article 11 CFR, are primarily addressed at the state or supranational organisations such as the EU; providers of intermediary services, as private parties, are not directly bound by article 11 CFR.[55] Providers of intermediary services may be subjected to a horizontal effect of freedom of expression, depending on the national jurisdiction,[56]

---

[48] European Court of Human Rights, 15 May 2023, *Sanchez v France*, 45581/15, paras 161-162; European Court of Human Rights, 16 June 2015, *Delfi v Estonia*, 64569/09.

[49] European Court of Human Rights, 13 February 2003, *Cetin et al v Turkey*, 40153/98 & 40160/98, para 64.

[50] European Court of Human Rights, 9 April 2019, *Navalnyy v Russia (no 2)*, 43734/14, para 80.

[51] Court of Justice of the European Union, 27 March 2014, *UPC Telekabel Wien v Constantin Film Verleih & Wega Filmproduktionsgesellschaft*, C-314/12, para 55-56.

[52] Karanjot Gill, 'Regulating Platform's Invisible Hand: Content Moderation Policies and Processes' (2021) 21 Wake Forest Journal of Business and Intellectual Property Law 171, 175–182; Eric Goldman, 'Content Moderation Remedies' (2021) 28 Michigan Technology Law Review 1. See on amplification and de-amplification: Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790; Tarleton Gillespie, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 Social Media + Society 205630512211175; Daphne Keller, 'Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard' [2021] Journal of Free Speech Law 227.

[53] See e.g. Robyn Caplan, 'Networked Platform Governance: The Construction of the Democratic Platform' (2023) 17 International Journal of Communication 3451.

[54] On how platforms govern freedom of expression on the internet, see e.g. Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598; Tarleton Gillespie, *Custodians of the Internet* (1st edn, Yale University Press 2018).

[55] De Gregorio (n 10) 196–200; Claudia E Haupt, 'The Horizontal Effect of Fundamental Rights' in Giovanni De Gregorio, Oreste Pollicino and Peggy Valcke (eds), *Oxford Handbook on Digital Constitutionalism* (Oxford University Press 2024) 10.

[56] Janneke Gerards, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023) 105., arguing that following Court of Justice of the European Union, 13 May 2014, *Google Spain v Mario Gonzalez*, C-131/32, it could be argued that freedom of information has some horizontal effects in EU law: "the obligation to take action to ensure the freedom of information or balance it against other

and can be affected by fundamental rights norms in the interpretation of private law, but such approaches are not clear-cut and differ per member state.[57] Similarly, the DSA, despite being inherently a market harmonisation instrument,[58] creates some fundamental rights obligations to providers of intermediary services.[59] Some scholars argue that the nature of individual fundamental rights is less compatible with the scale of content moderation,[60] and that application of fundamental rights to the digital realm in its current form requires rethinking.[61] Until such time, the realisation of fundamental rights in the digital sphere is dependent on platforms moderating content, and how regulators require platforms to include fundamental rights considerations in the content moderation process. DSCs affect freedom of expression against this background:[62] they enforce a regulation that directly affects the content moderation process, which invariably has effects on the freedom of expression of internet users. In their enforcement, they are required to take freedom of expression into account, both by application of the Charter and indirectly by the DSA.[63]

---

rights is, after all, directly imposed on the search engine operator." De Gregorio (n 13) also underlines that the Court of Justice of the European Union may be open to such an interpretation based on Court of Justice of the European Union, 27 March 2014, *UPC Telekabel Wien GmbH v. Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH*, C-314/12; Court of Justice of the European Union, 15 September 2016, *Tobias Mc Fadden v. Sony Music Entertainment Germany GmbH*, C-484/14. Also e.g. Amélie Heldt, *Intensivere Drittwirkung* (Mohr Siebeck 2023).

[57] See e.g. Bundesverfassungsgericht, 22 May 2019, *Der III. Weg*, 1 BvQ 42/19. More generally Aleksandra Kuczerawy and Joao Quintais, '"Must-Carry", Special Treatment and Freedom of Expression on Online Platforms: A European Story' (forthcoming, on file with author).

[58] Regulation 2022/2065, preamble; Treaty on the Functioning of the European Union, art. 114; Annegret Engel, Xavier Groussot and Gunnar Thor Petursson (eds), *New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights*, vol 13 (Springer Nature Switzerland 2025).

[59] See e.g. Regulation 2022/2065, art. 14(4): "Providers of intermediary services shall act in a diligent, objective and proportionate manner in applying and enforcing the restrictions referred to in paragraph 1, with due regard to the rights and legitimate interests of all parties involved, including the fundamental rights of the recipients of the service, such as the freedom of expressions, freedom of pluralism of the media, and other fundamental rights and freedoms as enshrined in the Charter." See also João Pedro Quintais, Naomi Appelman and Ronan Ó Fathaigh, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' [2023] German Law Journal 1.

[60] Evelyn Douek, 'The Limits of International Law in Content Moderation' (2020) 6 U.C. Irvine Journal of International, Transnational and Comparative Law 37; Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939.

[61] Gunther Teubner, 'Horizontal Effects of Constitutional Rights in the Internet: A Legal Case on the Digital Constitution' (2017) 3 The Italian Law Journal 193.

[62] Paradoxically despite the DSA not directly requiring them to engage with the legality of individual pieces of content, see Regulation 2022/2065, rec. 109: "[…] This Regulation does not require Member States to confer on Competent Authorities the task to adjudicate on the lawfulness of specific items of content."

[63] E.g. Regulation 2022/2065, art. 49(6); rec. 115

3. DSCs interacting with freedom of expression

DSCs, in their role as enforcer of the DSA, have a number of tasks that affect the freedom of expression of internet users. As iterated above, DSC tasks are predominantly aimed at content moderation on a systemic level, focusing on procedural safeguards and risk mitigation, not on individual pieces of content.[64] Their competence can be divided in two themes: powers that directly affect content moderation (II.3.a & II.3.b), and powers that affect content moderation indirectly through investigating or sanctioning intermediary service providers (II.3.c & II.3.d).

a. Powers that affect the content moderation process

DSC may affect content moderation through article 9 DSA, allowing 'relevant national judicial or administrative authorities' to order the provider of intermediary services to act against illegal content.[65] DSCs explicitly fulfil a role as go-between for judicial or administrative authorities issuing orders to remove content or issuing orders to provide information; they must be informed by- and inform other DSCs of the content of such orders.[66] The institutional set-up explained in Section II.1 creates a peculiarity in this framework. The DSC itself qualifies as a relevant national authority, if appointed to this extent under national administrative law.[67] Aside from the DSC, it is possible that the authority hosting the DSC, e.g. the consumer authority of a member state, qualifies as a relevant national administrative authority. In the first instance, the interface with freedom of expression by the DSC is clear: if the DSC issues an order to act against illegal content and the provider of intermediary services follows up on that order by limiting access to the litigious content, the right to freedom of expression of the internet user is affected. In the second instance, if the hosting authority orders to act against illegal content, the DSC may still be involved, and therefore interacting with the right to freedom of expression of internet users. DSCs are currently hosted at authorities having oversight tasks under national or EU law, which can qualify them as relevant national administrative authorities, allowing them to submit orders to act against illegal content or request information. This can cloud the oversight roles at one organisation. Ideally, oversight authorities separate their tasks: the requirement of independence for DSCs as laid down in article 50(2) DSA requires that "competent

[64] Wilman (n 3) 18. Regulation 2022/2065, rec. 109; *cf.* Rademacher in Hofmann and Raue (n 18) 870–872.

[65] Regulation 2022/2065, art. 9.

[66] Regulation 2022/2065, art. 9(3)(4) & 10(3)(4). Providing information to other DSCs reflects the wish to prevent *ne bis in idem* and contributes to the principle of mutual assistance *ex* art. 57(1).

[67] *Cf.* Regulation 2022/2065, rec. 109

authorities […] act in *complete independence* from private *and public* bodies, without the obligation or possibility to seek or receive instructions […]."[68] However, there is no guidance on what safeguards around the dual functioning of DSCs should look like in practice, meaning that, as said above, independence is 'rather thin'.[69] A consumer authority might issue an order to act against illegal content, but it would be unclear whether this is order originated from the DSC hosted at the consumer authority or the consumer authority itself. In any case, if the provider of intermediary services follows up on that order, the freedom of expression of the internet user is interfered with on behalf of the DSC or its hosting authority.

A second possible interference with freedom of expression is found in the explicit role for DSCs in formulating codes of conduct. Under article 45, DSCs should encourage and facilitate the creation of codes of conduct, together with the Commission.[70] In encouraging or facilitating, DSCs can take varying roles which may affect the content of codes of conduct, for example by encouraging moderation of a specific type of content.[71] Although Codes of Conduct are a voluntary effort, there is significant informal pressure exuding from encouragement or facilitation for providers of intermediary services to partake, which may lead to cynicism on the presentation of state-led regulatory efforts as 'voluntary'.[72] The 'voluntary' nature of such codes, which is reinforced by non-participation in Codes of Conduct being interpreted as a contra-indicator for DSA compliance,[73] can also be interpreted as an interference with freedom of expression of internet users that lacks a legal basis if platforms remove content on the basis of Codes of Conduct that are not codified in law.[74]

---

[68] Regulation 2022/2065, rec. 112. See also more detailed fn. 16.

[69] Spindler (n 17) para 36.

[70] Regulation 2022/2065, art. 45.

[71] Christina Angelopoulos and others, 'Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation' (Institute for Information Law 2016). Notable examples are The EU Code of Conduct on Countering Illegal Hate Speech Online, https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en#relatedlinks; The 2022 Strengthened Code of Practice on Disinformation, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

[72] Rachel Griffin, 'Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns' (2024) 2024 Technology and Regulation 167, 185.

[73] Regulation 2022/2065, rec. 104: "The refusal without proper explanations by a provider of an online platform or of an online search engine of the Commission's invitation to participate in the application of such a code of conduct could be taken into account, where relevant, whether the online platform or the online search engine has infringed the obligations laid down by this Regulation […]."

[74] Jacob van de Kerkhof, 'Jawboning Content Moderation from a European Perspective' in Charlotte van Oirsouw and others (eds), *European Yearbook of Constitutional Law 2023: Constitutional Law in the Digital Era* (TMC Asser Press 2024) <https://doi.org/10.1007/978-94-6265-647-5_4> accessed 13 December 2024.

This reasoning also applies to the enforcement of article 14 on terms and conditions and article 16 on notice and action mechanisms. Article 14 poses a number of requirements for intermediary service providers regarding their terms and conditions, such as providing information on policies, procedures, and tools used for content moderation, and stipulates that providers of intermediary services must act diligently, objectively and enforce their terms of services in a proportionate manner, with due regard for the freedom of expression and the pluralism of the media.[75] Terms and conditions are commonly used by platforms to form guidelines that shape speech on their platforms.[76] This is one of the first regulatory interventions in the drafting and enforcement of terms and conditions, and provides significant power for DSCs to intervene in private norm-setting, for example on moderating certain types of content under terms and conditions.[77] Quintais et al underline this operational concern and raise questions on the enforcement of article 14(4) in light of how well fundamental rights align with large-scale content moderation, as touched upon in II.2.[78] The enforcement of article can therefore have far-reaching consequences, since moderation on the basis of terms and conditions forms the overwhelming majority of all content moderation decisions.[79]

How DSCs enforce or sanction compliance with article 16 forms a similar intervention in the content moderation process. Article 16 requires intermediary service providers to facilitate the submission of notices on illegal content. The mechanisms by which notices are shaped can affect the degree to which users report illegal content.[80] More pertinent for DSCs effect on freedom of expression of users is the enforcement of article 16(6).[81] Article 16(6) requires that intermediary service providers process notices in a timely, diligent, non-arbitrary, and objective manner.[82]

---

[75] Regulation 2022/2065, art. 14(1); art. 14(4).

[76] Niva Elkin-Koren, Giovanni De Gregorio and Maayan Perer, 'Social Media as Contractual Networks: A Bottom Up Check on Content Moderation' (2022) 107 Iowa Law Review 987; João Pedro Quintais, Giovanni De Gregorio and João C Magalhães, 'How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications' (2023) 48 Computer Law & Security Review 105792.

[77] Quintais, Appelman and Ó Fathaigh (n 59) 29. See also Rademacher in Hofmann and Raue (n 18) 859.

[78] Quintais, Appelman and Ó Fathaigh (n 59).

[79] Rishabh Kaushal and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024).

[80] See in the context of NetzDG: Ben Wagner and others, 'Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <https://dl.acm.org/doi/10.1145/3351095.3372856> accessed 18 October 2023; Amélie Heldt, 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports' (2019) 8 Internet Policy Review.

[81] Regulation 2022/2065, art. 16(6).

[82] See Daniel Holznagel, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (2024) 25 Computer Law Review International 172.; Objectivity can be contentious: currently X is under investigation by the Commission *inter alia* for not handling notices objectively. See

The standards by which intermediary service providers moderate on the basis of notices determines the scope of freedom of expression of users whose content notices are submitted on. Enforcement aimed at the processing of such notices therefore shapes the right to freedom of expression online.

In a further intervention in the content moderation process, DSCs are tasked with certifying stakeholders in the content moderation process such as out-of-court dispute settlement bodies and trusted flaggers. If a user disagrees with the decision of an internal complaint-handling system on a complaint regarding a content moderation decision,[83] they have the right under the DSA to select any certified out-of-court dispute settlement body to appeal that decision.[84] Out-of-court dispute settlement bodies are independent entities certified by the DSC that decide on such matters in a fair, swift and cost-effective manner.[85] A well-known current example of such an entity is the Meta Oversight Board, which, although it has been criticized as a form of 'governance-washing', has a significant norm-setting power in issuing decisions, affecting Meta's content moderation practices.[86]

Out-of-court dispute settlement bodies under article 21 may have a less profound impact due to the fact that their decisions are non-binding.[87] Their impact is also limited by the fact that internet users are free to choose the out-of-court dispute settlement body of their preference; users are free to avoid out-of-court dispute settlement bodies that do not align with their values.[88] However, the normative power that non-binding decisions exude over content moderation practices can shift existing policy in favour of avoiding future adverse decisions, in line with the reasoning of the 'chilling effect' platform liability can have on freedom of expression.[89] Authors so far have different visions on the impact out-of-court dispute

---

https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709; see further Matteo Fabbri, 'The Role of Requests for Information in Governing Digital Platforms Under the Digital Services Act: The Case of X' (2025) 6 Journalism and Media 41.

[83] Regulation 2022/2065, art. 20.

[84] Regulation 2022/2065, art. 21(1).

[85] Regulation 2022/2065, art. 21(3)(e).

[86] Lucas Henrique Muniz Da Conceicao, 'A Constitutional Reflector? Assessing Societal and Digital Constitutionalism in Meta's Oversight Board' [2024] Global Constitutionalism 1; Kate Klonick, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 Yale Law Journal 2418; David Wong and Luciano Floridi, 'Meta's Oversight Board: A Review and Critical Assessment' (2023) 33 Minds and Machines 261.

[87] Regulation 2022/2065, art. 21(2). See also Ruairí Harrison, Jonny Shipp and Aebha Curtis, 'Settling DSA-Related Disputes Outside the Courtroom: The Opportunities and Challenges Presented by Article 21 of the Digital Services Act' (2024).

[88] Regulation 2022/2065, art. 21(1). Currently, the number of out-of-court dispute settlement bodies that are onboarded with online platforms is still somewhat limited, however.

[89] Court of Justice of the European Union, 6 October 2020, *La Quadrature du Net*, C-511/18, C-512/18 &C-520/18, para 113; Court of Justice of the European Union, 8 April 2014, *Digital Rights Ireland*, C-293/12 & C-594/12, para 25; E.g. Aleksandra Kuczerawy, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (2015) 31

settlement bodies will have on content moderation and freedom of expression in general: Wimmers argues that certifying more entities to decide on content moderation leads to further fragmentation, which goes against the spirit of the DSA.[90] He further argues that "free speech disputes do not lend themselves to settlement by private bodies."[91] Ortolani proposes a more nuanced 'wait-and-see' approach with regards to out-of-court dispute settlement bodies.[92] Assuming that out-of-court dispute settlement bodies have a significant impact on content moderation because of the normative weight of their decisions recalibrating content moderation practices, a question rises about the DSC certifying such bodies: they have the power to accredit entities that affect content moderation, an impact that is felt Union-wide, making the selection and accreditation of these entities an important interface with user's freedom of expression. The role of individual DSCs in this is crucial: if an out-of-court dispute settlement body fails to meet the requirements laid down in article 21(3), only the certifying DSC can revoke that bodies' status, limiting the safeguards for EU citizens outside that Member State that may be affected by the normative power exuding from an out-of-court dispute settlement bodies' decisions.

The selection and certification of trusted flaggers presents a similar interface with freedom of expression. Trusted flaggers are state- and non-state entities that have priority access to the content moderation process, meaning that notices, 'flags',[93] they submit with the provider of intermediary services are treated expediently and with priority.[94] DSCs appoint trusted flaggers based on expertise, independence, diligence, accuracy and objectivity; priority of those flags is granted based on the experience such entities hold on a specific topic, usually leading to higher quality

Computer Law & Security Review 46; Giancarlo Frosio and Christophe Geiger, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 European Law Journal 31; Gerards (n 56) 103.

[90] Jörg Wimmers, 'The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act' (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 381.

[91] ibid 383. This critique applies to freedom of expression in content moderation in general; see Douek (n 60); Sander (n 60); Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 University of California, Davis Law Review 1149.

[92] Pietro Ortolani, 'If You Build It, They Will Come' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-build-it/> accessed 30 September 2024.

[93] Kate Crawford and Tarleton Gillespie, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 New Media & Society 410.

[94] Regulation 2022/2065, art. 22; see also Naomi Appelman and Paddy Leerssen, 'On Trusted Flaggers' (2022) 24 Yale Journal of Law & Technology 452; Sebastian Felix Schwemer, 'Trusted Notifiers and the Privatization of Online Enforcement' (2019) 35 Computer Law & Security Review 105339.

notices aiding online platforms in their content moderation process.[95] The status of trusted flagger can, especially when state entities act as trusted flaggers, exude pressure on providers of intermediary services to comply with removal requests.[96] This has a direct impact on the visibility of user-generated content. Trusted flaggers can report content on the basis of national or EU law following article 3(h), creating a risk for extraterritorial application of national law.[97] The fact that users do not have the freedom to select the trusted flaggers involved in their content moderation process exacerbates these concerns as opposed to the freedom they have in choosing out-of-court dispute settlement bodies. Selecting what entity takes part in the content moderation process requires DSCs to take into account the freedom of expression of internet users, or at least consider the capacity of the trusted flagger to do so. However, similar to article 21(3), the DSA does not provide clear guidance in this regard on article 22(2), and the Commission is yet to exercise its power in issuing guidelines on the certification of trusted flaggers under article 22(8).[98]

b. The European Board of Digital Services

DSCs cooperate across the EU in the European Board of Digital Services. The Board is tasked with harmonising DSA enforcement across the EU. On a general level, DSCs contribute in the Board to drafting guidelines on matters covered by the DSA.[99] Although the Commission has explicit competences to issue guidelines, its role in the Board is limited because it has no voting powers. DSCs, in their norm-setting role in the Board, influence how the DSA is enforced throughout the European Union. In this role, they need to cooperate with 26 other DSCs, but larger DSCs, hosting more online platforms, may have more influence in the setting of guidelines, since they oversee more and larger providers of intermediary services and have more resources to engage in norm-development.[100] They have, as such, a more prominent position in shaping freedom of expression online than other DSCs, a 'Dublin-Effect'.[101]

---

[95] Regulation 2022/2065, art. 22(2). See also Benjamin Raue, 'Artikel 22 Vertrauenswürdige Hinweigeber' in Franz Hofmann and others (eds), *Digital Services Act: Gesetz über digitale Dienste* (1. Auflage, Nomos 2023).

[96] Jacob van de Kerkhof, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' in José van Dijck and others (eds), *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values* (Amsterdam University Press 2025).

[97] Regulation 2022/2065, art. 16(1) & art. 3(h).

[98] One could argue that fundamental rights expertise is captured in "particular expertise and competence for the purpose of detecting, identifying and notifying illegal content" in art. 22(2), but this is not explicitly mentioned.

[99] Regulation 2022/2065, art. 61(2)(a)&(b).

[100] See also Jennifer Orlando-Salling, 'The Digital Services Act in the European Periphery: Critical Perspectives on EU Digital Regulation' [2025] European Law Open 1. taking a European core – periphery perspective on this matter.

[101] A play on Bradford's seminal 'Brussels effect', see Anu Bradford, 'The Brussels Effect' (2012) 107 Northwestern University Law Review 1. See also in commenting on the Draft DSA: Daniel Holznagel, 'Ireland Cannot Do It Alone' (*Verfassungsblog*, 27 April 2021) <https://verfassungsblog.de/dsa-ireland/> accessed 23 April 2025.

Aside from general guidelines on DSA enforcement that may affect content moderation, there is an explicit role for the Board in crisis situations. The Board makes recommendations to the Commission on triggering the crisis response mechanism laid down in article 36(1). In such cases, VLOPSEs must enact crisis response mechanisms that include the assessment of whether the functioning of their services contributes to a serious threat to public security or public health.[102] VLOPSEs must then apply specific, effective, and proportionate measures to mitigate threats and report back on these measures with the Commission.[103] These measures can have a significant impact on content moderation in the European Union: case law of the European Court of Human Rights such as *Cengiz et al v Turkey* and *Yilderim v Turkey* underline that crises can interfere with freedom of expression on the internet.[104] Similarly, the COVID-19-crisis showed that internet services are key sources of information on public health, sparking a rich debate on content moderation, freedom of expression, and disinformation.[105] The power to issue recommendations on enacting crisis protocols therefore has a significant impact on the freedom of expression of internet users. The influence of a single DSC in the grander scheme of article 36 may be limited: a recommendation is issued by the Board, not by a single DSC, and the Commission has the final say in declaring states of crisis.[106] However, if a single DSC were particularly influential due to their size and the number of providers of intermediary services they host (the 'Dublin-effect'), they may inspire the Commission to declare a state of crisis. It may be difficult for smaller, less influential DSCs to acquire a similar status. In the recent case of the alleged interference of the Romanian elections via TikTok therein, it was difficult for the Romanian DSC to trigger a similar crisis situation for TikTok.[107]

[102] Regulation 2022/2065, art. 36(1).

[103] Regulation 2022/2065, art. 36(3)(a).

[104] European Court of Human Rights, decision of 1 March 2016, *Cengiz et al v Turkey*, 48226/10 & 14027/11; European Court of Human Rights, decision of 18 March 2013, *Ahmet Yilderim v Turkey*, 3111/10.

[105] E.g. Michael Karanicolas, 'Even in a Pandemic, Sunlight Is the Best Disinfectant: COVID-19 and Global Freedom of Expression' (2021) 22 Oregon Review of International Law 1; Berdien BE van der Donk, 'Should Critique on Governmental Policy Regarding Covid-19 Be Tolerated on Online Platforms? An Analysis of Recent Case-Law in the Netherlands' (2021) 13 Journal of Human Rights Practice 426.

[106] Regulation 2022/2065, art. 36(1): "Where a crisis occurs, the Commission, acting upon a recommendation of the Board, may adopt a decision […]; *cf.* art. 56(2).

[107] Of course, this is difficult since the alleged misinformation campaign was largely only discovered after the fact, being that the election had already transpired. The Commission has since released an 'election toolkit', see: <https://digital-strategy.ec.europa.eu/en/news/commission-presents-new-best-practice-election-toolkit-digital-services-act> accessed 17 June 2025.

c. Powers related to investigating service providers

To enforce the DSA, DSCs are attributed a range of investigatory and sanctioning powers to relating to providers of intermediary services for non-compliance with the DSA. Part of these powers is aimed at creating knowledge about platform practices. Regulators suffer from a significant knowledge deficit, and the DSA provides a toolbox for overcoming that. [108] DSCs can request updated transparency reports from intermediary service providers,[109] and request access with the Commission to risk assessments of VLOPSEs established in its jurisdiction.[110] DSCs can also request access to data with VLOPSEs under article 40(1) for the purpose of monitoring and assessing compliance with the DSA with a reasoned request.[111] DSCs vet researchers, meaning they issue data access requests on behalf of researchers seeking to study systemic risks (article 34(1)) under article 40(8).[112] In granting applications, DSCs have a role in shaping the interpretation of systemic risk, even though that falls under the Commissions competence.[113] The information gathered through these research and data access requests shapes the understanding and therefore the enforcement of the systemic risk assessment, mitigation, and auditing framework.[114]

Aside from requesting data access, it can request information from providers of intermediary services, as well as auditing organisations, including all members of staff.[115] It can further inspect the premises of providers of intermediary services and confiscate information relating to suspected infringement.[116] These are so-called 'dawn-raids', which are already prevalent in other fields of law, such as consumer protection and competition law. In those fields they are considered impactful

[108] E.g. Regulation 2022/2065, rec. 96.

[109] Regulation 2022/2065, art. 24(3)&(4).

[110] Regulation 2022/2065, art. 34(3).

[111] Regulation 2022/2065, art. 40(1).

[112] Regulation 2022/2065, art. 40(8), see also Draft Commission Delegated Regulation supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data pursuant to Article 40 of Regulation (EU) 2022/2065; see further Anna Liesenfeld, 'The Legal Significance of Independent Research Based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act' [2024] European Journal of Risk Regulation 1.

[113] Regulation 2022/2065, art. 56(2); on the cooperation with the Commission in issuing guidelines to mitigate systemic risks, see art. 35(3).

[114] Mateus Correia de Carvalho, 'It Will Be What We Want It to Be: Sociotechnical and Contested Systemic Risk at the Core of the EU's Regulation of Platforms' AI Systems' [2025] Journal of Intellectual Property, Information Technology and Electronic Commerce Law 35.

[115] Regulation 2022/2065, art. 51(1)(a).

[116] Regulation 2022/2065, art. 51(1)(b).

measures from a procedural- and human rights perspective, underlining the DSCs significant investigatory powers in the DSA.[117]

d. Powers related to sanctioning intermediary service providers

If investigatory powers indicate a violation of the DSA, the DSC can use its sanctioning powers to combat infringement. It can agree on binding commitments with providers aimed at ceasing the infringement,[118] or order cessation of the infringement under a periodic penalty.[119] It can also fine providers of intermediary services[120] or adopt interim measures that avoid the risk of serious harm arising from the infringement.[121] Penalties imposed on providers of intermediary services must be effective, proportionate and dissuasive; if sanctions do not lead to cessation of the infringement and the infringement causes serious harm that cannot be avoided, the DSC has the power to require the management body of the provider of intermediary services to examine the situation and adopt an action plan that must be carried out and reported upon.[122] If the infringement continues and entails a criminal offence and continues to involve a threat to the life or safety of people, the DSC can request a competent judicial authority in its Member State to temporarily restrict access to the service infringing on the DSA, or, if such is technically unfeasible, to the online interface of the provider of intermediary services.[123] The power to block access to platforms has been linked to freedom of expression infringements by the European Court of Human Rights, and underlines the enforcement power that DSCs hold.[124]

e. Safeguards around DSA enforcement

The DSA must be enforced in light of the principle of proportionality: any action taken must be proportional to the nature, gravity, duration, recurrence of the infringement, without unduly restricting access to the service; measures must also take into account the technical- and operational capacity of the provider of the

[117] Helene Andersson, *Dawn Raids Under Challenge: Due Process Aspects on the European Commission's Dawn Raid Practices* (Bloomsbury Publishing 2024). See also Court of Justice of the European Union, judgment of 20 June 2018, *Cesky Drahy v Commission,* T-325/16; European Court of Human Rights, judgment of 21 March 2017, *Janssen Cilag S.A.S. v France,* 33931/12. In a GDPR context, the power to 'dawn-raid' is also provided in article 58(1)(f) Regulation 2016/679, but it has not received similar criticism.

[118] Regulation 2022/2065, art. 51(2)(a).

[119] Regulation 2022/2065, art. 51(2)(d).

[120] Regulation 2022/2065, art. 51(2)(c).

[121] Regulation 2022/2065, art. 51(3).

[122] Regulation 2022/2065, art. 51(3)(a).

[123] Regulation 2022/2065, art. 51(3)(a).

[124] E.g. European Court of Human Rights, decision of 18 March 2013, *Ahmet Yilderim v Turkey*, 3111/10.

intermediary service concerned.[125] Furthermore, Member States need to lay down rules and procedures to ensure that the exercise of sanctioning powers is compliant with applicable national law, the Charter, and general principles of EU law.[126] Affected parties of DSA enforcement enjoy the right to an effective judicial remedy through a range of private- and public enforcement mechanisms.[127] Public enforcement mechanisms involve filing complaints with Digital Services Coordinators under article 53 or access to national courts.[128]

In terms of substantive fundamental rights, Article 51(6) emphasises the right to respect for private life and the rights of defence, including the right to be heard and access to the file.[129] Recital 153 further emphasises that the right to freedom of expression must also be taken into account in DSA enforcement.[130] Exactly how the right to freedom of expression is taken into account in the enforcement of the DSA is not clarified. DSCs likely have fundamental rights anchored in their national system as well, but scope of those fundamental rights may differ, as illustrated in Section II. This raises the question whether freedom of expression is sufficiently anchored in DSA enforcement.

4. Interim Conclusion

Section I has outlined the various interfaces that the DSC's role under the DSA has with the freedom of expression of internet users. DSCs hold significant power that

---

[125] Treaty on the European Union, art. 5(4); Regulation 2022/2065, art. 51(5). In this regard, the investigatory powers outlined in the previous subsection have a key role: in platform regulation, and intermediary service providers often claim that interventions requested by administrative or judicial authorities are technically unfeasible and therefore disproportionate. Examples of such arguments are found in: Court of Justice of the European Union, judgment of 13 May 2014, *Google Spain v AEPD & Mario Gonzalez*, C-131/12; Court of Justice of the European Union, judgment of 3 October 2019, *Eva Glawischnig-Piesczek v Facebook Ireland Ltd*, C-18/18; Court of Justice of the European Union, judgment of 22 June 2021, *Peterson v Google and Elsevier v Cyando*, C-682/18 and C-683/18; Court of Justice of the European Union, judgment of 26 April 2022, *Republic of Poland v European Parliament & Council of the European Union*, C-401/19. When it comes to the technical feasibility of content moderation orders, there is an awkward tension between the fact that Member States may not order providers of intermediary services to adopt general monitoring instruments under Article 8 DSA, but still request that *de facto* they employ general monitoring instruments in order to adhere to the highest industry standards (for example Article 19(4)(b) Directive 2019/790, to avoid liability for copyright infringement). Such is also established in European Court of Human Rights, 5 September 2023, *Zoechling v Austria,* 4222/18, para 13; European Court of Human Rights, 15 May 2023, *Sanchez v France,* 45581/15, para 190. See further on this Martin Senftleben and Christina Angelopoulos, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the e-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' [2020] Amsterdam/Cambridge.

[126] Regulation 2022/2065, art. 51(6), see also Charter of Fundamental Rights of the European Union, art. 52(1)

[127] Charter of Fundamental Rights of the European Union, art. 47.

[128] Regulation 2022/2065, art. 53; rec. 39; see also Giovanni De Gregorio and Simona Demková, 'The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe' in Charlotte Van Oirsouw and others (eds), *European Yearbook of Constitutional Law 2023*, vol 5 (TMC Asser Press 2024).

[129] Charter of Fundamental Rights of the European Union, art. 8 & 48.

[130] Regulation 2022/2065, rec. 153.

intervene with the content moderation process, and as such the DSC has the potential to influence freedom of expression on the internet: a responsibility that requires considerable safeguards. These safeguards, outlined in II.3.e, are relatively limited. Under the DSA, the DSC also has the responsibility to protect the freedom of expression of internet users. In that protecting role, the decentralised network of enforcement of the DSA presents a challenge. This is explored in the next section.

III. Decentralised enforcement of the DSA leading to different applications of the DSA

This section builds on the previous section by explaining how the decentralised enforcement framework in the DSA can fragment its application, which can diminish the protection of freedom of expression. Decentralised enforcement has benefits: enforcement takes place in close proximity to where the effects are felt, builds on 'political and cultural know-how', and Member States maintain administrative and procedural autonomy.[131] However, this article proposes that the decentralised structure of the DSA can lead to different levels of protection of freedom of expression, which is not sufficiently mitigated in DSA enforcement network. This is explained following an introduction of the DSA's decentralised enforcement structure.

1. The DSA's decentralised enforcement structure

The DSA's enforcement is dictated by national administrative authorities and the European Commission, relying on the competence structure described in Section II.1.[132] Rademacher and Marsch describe the DSA enforcement network as both horizontally and vertically centralised.[133] DSA enforcement is centralised horizontally by a country-of-origin principle,[134] somewhat resembling the one-stop-shop found in the General Data Protection Regulation ('GDPR').[135] That framework can be used as a frame of reference for decentralised enforcement; other fields of EU

[131] Court of Justice of the European Union, 16 December 1976, *Rewe v Landwirtschaftskammer für das Saarland*, C-33/76; Schütze (n 13).

[132] ibid; Katarzyna Jancewicz, Pieter van Cleynenbreugel and Julien Bois, 'Inconsistent Administrative Enforcement of EU Law at Member State Level: The Lisbon Treaty's Hidden Constitutional Challenge' (2024) <https://orbi.uliege.be/bitstream/2268/314238/1/EUDAIMONIA_WP20242.pdf> accessed 17 October 2024.

[133] Rademacher & Marsch in Hofmann and Raue (n 18) 973–974.

[134] Regulation 2022/2065, art. 56(1).

[135] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), *OJ L 119, 4.5.2016*, p. 1-88, art. 55 & 56.

law, such as consumer or competition law, have a different competence structure.[136] DSCs have competence over the providers of intermediary services located in their jurisdiction, but if intermediary service providers fall outside of their competence, they must rely on other DSCs to investigate and enforce.

The DSA creates a legally binding cooperation mechanism that is distinct from the framework in the GDPR. Under article 58(1), a DSC of destination (i.e. in which the intermediary service is delivered) can request the DSC of establishment to take investigatory and enforcement measures to ensure compliance with the DSA.[137] These requests require the DSC of establishment to provide an assessment of the suspected infringement and an explanation of any investigatory or enforcement measures taken within two months following the request.[138] If deemed insufficient, the Commission acts as a dispute resolution mechanism, issuing non-binding conclusions to the DSC of establishment indicating whether the investigatory or enforcement measures are sufficient.[139] Similarly, the DSC of establishment can request, or be recommended to, launch a joint investigation, enabling more enforcement capacity over complicated matters.[140] These avenues help overcome potentially inactive administrative authorities; without such cooperation mechanisms, activation would have to be realised through infringement procedures under article 258 TFEU.

Decentralised enforcement structures are no novelty in EU law. However, DSA enforcement is also vertically centralised, which is markedly different from other fields of law. For example, in data protection law the Commission is a relatively subordinate actor due to the strict independence of national data protection authorities,[141] and in consumer protection law, the Commission does not have significant enforcement powers, but is tasked with streamlining enforcement by independent national authorities.[142] As discussed in Section II.1, the Commission has

[136] For example, consumer law requires the coordinated enforcement with a network of consumer protection authorities under Regulation 2017/2394, see also Catalina Goanta and Jerry Spanakis, 'Discussing The Legitimacy of Digital Market Surveillance' [2022] Stanford Journal of Computational Antitrust 44. Competition law is marked by the fact that the European Commission is the only enforcer in cases of merger control (Regulation 2004/139) or the DMA (Regulation 2022/1925), and enforces concurrently with national competition authorities in cases concerning art. 101 TFEU and 102 TFEU.

[137] Regulation 2022/2065, art. 58(1).

[138] Regulation 2022/2065, art. 58(5).

[139] Regulation 2022/2065, art. 59(3).

[140] Regulation 2022/2065, art. 60(1).

[141] Regulation 2016/679, art. 51(1), rec. 117. Treaty on the Functioning of the European Union, art. 16(2); Charter of Fundamental Rights of the European Union, art. 8. See also Lisette Mustert, 'Cross-Border Enforcement of the GDPR by Independent Administrative Authorities' (University of Luxembourg 2023) 54–63.

[142] Regulation 2017/2394, see also Goanta and Spanakis (n 136).

exclusive competence over VLOPSE-specific obligations under Section 5 of Chapter 3, and can initiate proceedings against VLOPSEs for other obligations if appropriate. The Commission's position raises concerns about the independence requirement for enforcement of the DSA, since the Commission is an inherently political institution.[143] Adding to this, the Commission takes significant powers regarding fundamental rights of EU citizens, despite the DSA being entirely based on a legal basis of market harmonisation under article 114 TFEU, creating the concern for competence creep.[144] From a practical perspective however, the Commission's involvement alleviates the enforcement burden on national coordinators for dealing with the largest intermediary service providers, and prevents regulatory capture of a single national authority as it occurred in data protection.[145] The vertical centralisation of enforcement in the DSA also mitigates concerns arising from a one-stop-shop model; however three factors troubling freedom of expression in the DSA's enforcement remain: (i) different scope of freedom of expression being facilitated in the DSA (ii) differences in the procedural frameworks in member states, leading to different levels of protection, and (iii) differences at an institutional level – the DSA is enforced by different types of institutions with varying levels of enforcement experience and resources, which can affect the nature of enforcement and its effectiveness.

2. Different scope of freedom of expression facilitated by the DSA

Fragmentation in the application of the DSA may be caused by substantively different interpretations of its contents by Digital Services Coordinators.[146] The primary concern is the wide definition of 'illegal content' in article 3(h), a core concept throughout the DSA. In its current phrasing, article 3(h) is a content-agnostic enigma:[147] "any information that […] is not in compliance with Union law, or the law

[143] Rademacher in Hofmann and Raue (n 18) 851.

[144] Engel, Groussot and Petursson (n 58) 24.

[145] Rademacher in Hofmann and Raue (n 18) 991.

[146] It could be argued that different interpretations of procedural norms in the DSA can have an effect on the freedom of expression of internet users, for example affecting their right to an effective remedy. This is beyond the scope of this article.

[147] It is the intention of article 3(h) to provide a wide definition. Regulation 2022/2065, rec 12: "[…] the concept of 'illegal content' should broadly reflect the existing rules in the offline environment. In particular, the concept of 'illegal content' should be defined broadly to cover information relating to illegal content, products, services and activities. […] The question is whether this definition can lead to competency creep: a situation in which the Digital Services Coordinator, building on a broad definition of its subject matter (see, by comparison with data protection law, Purtova's commentary on how the broad concept of personal data in the GDPR leads to an overly wide scope of application: Nadezhda Purtova, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40.) can assume competence over cases that could also fall under other enforcement authorities, particularly in light of the overlap between GDPR (Regulation 2016/679), DSA (Regulation 2022/2065), and AI Act (Regulation 2024/1689, with the exception of systemic risk assessments following rec. 118). See on this Paul de Hert and Paweł Hajduk, 'EU Cross-Regime Enforcement, Redundancy

of any Member State which is in compliance with Union law."[148] The definition is central in the DSA and weighs upon aspects that the DSC needs to enforce. In theory, its scope of application is near-limitless; Mauritz explains that because article 3(h) obfuscates the question what law is applicable, because it lacks a clear connecting factor connecting circumstances of the case to applicable national law.[149] Holznagel underlines this with an example:[150] holocaust denial is explicitly criminalised in some countries in the EU, such as Germany and France, but not in other countries, where it may be covered under more general hate speech provisions.[151] Another example: currently Hungary's anti-LGBT laws outlaw companies advertising in solidarity with the LGBT-community.[152] This begs the question whether, for example, the Irish DSC must sanction intermediary service providers that fail to suspend accounts spreading content related to these examples under article 23(1).[153]

The only back-stop delimiting 'illegal content' in the DSA is found in the phrasing: "in compliance with Union law", meaning that the Charter can render any law incompatible with EU law inapplicable. This operationalises the principle of mutual trust[154] and confirms a basic understanding on what defines illegal content. This appears a sufficient safeguard, but in practice it creates a number of tensions. Firstly, it forces entities seeking a definition for 'illegal content', such as providers of

---

and Interdependence. Addressing Overlap of Enforcement Structures in the Digital Sphere after Meta.' (2024) 2024 Technology and Regulation 291, commenting on Court of Justice, 4 July 2023, *Meta Platforms Inc v Bundeskartellamt*, C-252/21, presenting a similar struggle between competition- and data protection authorities. See also Simona Demková and Giovanni De Gregorio, 'The Looming Enforcement Crisis in European Digital Policy' (*Verfassungsblog*, 10 February 2025) <https://verfassungsblog.de/the-looming-enforcement-crisis-ai-dsa-eu/> accessed 25 April 2025.

[148] Regulation 2022/2065, art. 3(h).

[149] Franziska Katharina Mauritz, 'To Define Is Just to Define' (*Verfassungsblog*, 8 April 2024) <https://verfassungsblog.de/to-define-is-just-to-define/> accessed 20 August 2024.

[150] Daniel Holznagel, 'Safeguarding Adequate Rights Enforcement through the Digital Services Act (DSA)' <https://hateaid.org/wp-content/uploads/2022/05/2021-09-08-DSA-How-to-improve-enforcement-through-the-DSA.docx.pdf> accessed 26 August 2024.

[151] See on this: P Lobba, 'Holocaust Denial before the European Court of Human Rights: Evolution of an Exceptional Regime' (2015) 26 European Journal of International Law 237: "the utterance alone does not warrant criminal punishment." See also for example: Council Framework Decision 2008/913/JHA of 28 November 2008 on certain forms and expressions of racism and xenophobia by means of criminal law. Generally holocaust denial is not covered by the right to freedom of expression as laid down in article 10 ECHR: European Court of Human Rights, 23 September 1998, *Lehideux & Isorni v France,* 55/1997/839/1045; also Oversight Board, 2023, *Holocaust Denial*, 2023-022-IG-UA.

[152] Act LXXIX of 2021 on taking more severe action against paedophile offenders and amending certain acts for the protection of children; See for commentary on its threat to the EU legal order: Kim Lane Scheppele, 'The Treaties Without a Guardian: The European Commission and the Rule of Law' (2023) 29 Columbia Journal of European Law 93; Beatriz Botero Arcila and Rachel Griffin, 'Social Media Platforms and Challenges for Democracy, Rule of Law and Fundamental Rights' (European Parliament 2023) PE 743.400.

[153] Regulation 2022/2065, art. 23(1).

[154] Court of Justice of the European Union, 18 December 2014, Opinion 2/13, para 168; Treaty on the European Union, art. 2.

intermediary services, trusted flaggers, out-of-court dispute settlement bodies, but also Digital Services Coordinators, to evaluate whether the law they seek to apply is compliant with the Charter. This is a complicated endeavour that, aside from raising a range of legitimacy- and rule of law concerns related to this quasi-constitutional review of the litigious set of laws, could exceed the expertise of such entities, therefore proving an ineffective backstop. Secondly, it is unlikely that the interpretation of this backstop is uniform, potentially endangering the coherence of Union law. Article 61(2) attributes a harmonising role in this matter to the Board of DSCs.[155] However, this does not provide a clear avenue for DSCs seeking a guiding interpretation of the DSA with the Board, and even less so for out-of-court dispute settlement bodies and trusted flaggers, who ultimately interpret article 3(h) by themselves. This may result in content being removed on the basis of Charter-incompatible or inapplicable law, contravening the legality requirement of interference with the right to freedom of expression laid down in article 52(1) CFR.[156]

3. Different procedures on a national level

Differences in the implementing law of the DSA can also affect its application. This occurs when, under the discretion that member states have to appoint supervisory authorities for each of the DSA's oversight tasks, different entities across member states can have the same supervisory tasks under the DSA.[157] The degree to which these enforcement structures can affect enforcement can be illustrated using the Netherlands and Austria as an example. An internet user in Austria complains[158] with the Austrian DSC (*KommAustria*) on being presented with advertisements based on profiling as defined in article 4(4) GDPR,[159] by the provider of an online platform with its main establishment in the Netherlands, which infringes on article 26(3) DSA.[160] The Austrian DSC must refer this matter to the Dutch DSC (*Autoriteit Consument en Markt)*.[161] The Dutch DSC will under the Dutch implementation law forward the case to the Dutch Data Protection Authority ('DPA'), who is attributed the task to oversee article 26(3).[162] The Dutch DPA can then fine the intermediary

[155] Regulation 2022/2065, art. 61(2)(b).

[156] See also European Court of Human Rights, 26 April 1979, *Sunday Times v the United Kingdom,* 6538/74, para 49.

[157] Regulation 2022/2065, art. 49(2), rec. 109.

[158] Regulation 2022/2065, art. 53; *Bundesgesetz über den Koordinator für digitale Dienste nach der Verordnung(EU) 2022/2065 über einen Binnenmarkt für digitale Dienste,* StF: BGBI. I Nr. 182/2023 (NR: GP XXVII RV 2309 AB 2344 S. 247 S. 247 BR: 11366 AB 11400 S. 961, para 2(3)11.

[159] Regulation 2016/679, art. 4(4).

[160] Regulation 2022/2065, art. 26(3).

[161] Regulation 2022/2065, art. 57(2) & 58(1).

[162] *Uitvoeringswet digitaledienstenverordening,* art. 3.2 para 1.

service provider or make binding commitments, also under Dutch administrative law.[163] In the reverse scenario, if the user were a Dutch citizen complaining about an online platform established in Austria, the Dutch DSC would forward the case to the Dutch DPA.[164] The Dutch DPA would forward the case to the Austrian DPA, who then forwards the case to *KommAustria*, because *KommAustria* is the competent authority to enforce the DSA, including article 26(3) under the Austrian implementation.[165] Final result is that, depending on the national implementation of the DSA, the same complaint is addressed by two different institutions.

The enforcement structure illustrated above is complex, raising concerns that are only somewhat mitigated in the DSA. The reality is that this enforcement structure can be further complicated if composite procedures are added because the member state has a federal structure, meaning that competences may be attributed to regional rather than national authorities. In other fields of law, complex enforcement structures have been shown to ultimately lead to reduced accountability and protection of individual rights.[166] In DSA enforcement, each coordinating DSC is responsible for the uninterrupted functioning of their national composite procedures.[167] However, information deficits due to insufficient sharing of information;[168] supervision of administrative action; and, in the case of one-stop-shop mechanisms, administrative paralysis due to the overload that can arise from many platforms being located in a single member state[169] may interrupt composite procedures. Similarly, complaints under article 53 suffer reduced potential for judicial review if they involve a cross-border aspect, since in principle only courts of the same legal order are able to review decisions of administrative authorities. This implies that complainants in a different jurisdiction may not be able to challenge a decision taken by a foreign DSC if it is adopted by their own DSC in their own jurisdiction, or that such a challenge requires the national court of one jurisdiction to

[163] *Algemene wet bestuursrecht*, art. 5:20.

[164] Uitvoeringswet Digitaledienstenverordening, art. 3.2(1).

[165] *Bundesgesetz K-DDG*, para 2(1). The Austrian DSC must interact with other competent authorities, and have regular exchanges of views on issues that are concerned with the regulation of providers of intermediary services, and thus take the Austrian DPA's view into account.

[166] Giulia Gentile and Orla Lynskey, 'Deficient by Design? The Transnational Enforcement of the GDPR' (2022) 71 International and Comparative Law Quarterly 799.

[167] Rademacher in Hofmann and Raue (n 18) 850.

[168] Herwig CH Hofmann, 'Composite Decision Making Procedures in EU Administrative Law' in Herwig CH Hofmann and Alexander H Türk (eds), *Legal Challenges in EU Administrative Law* (Edward Elgar Publishing 2009) 3–4.

[169] 'Europe's Enforcement Paralysis' (Irish Council for Civil Liberties 2021) <https://www.iccl.ie/wp-content/uploads/2021/09/Europes-enforcement-paralysis-2021-ICCL-report-on-GDPR-enforcement.pdf> accessed 24 October 2024.

review the assessment of that of another court in a different jurisdiction. Seeing as the national law applied in these cases is crucial (see II.2), this may leave complainants with limited potential for judicial review.[170]

Administrative law can also disperse levels of protection under the DSA. DSCs are ultimately rooted in their respective national legal orders, creating an interaction between national and EU law. Rules for such interactions are missing, according to van Cleynenbreugel.[171] National authorities could afford more protection under national administrative law than required under principles of EU administrative law,[172] resulting in different levels of protection, which may subsequently affect the right to freedom of expression. National administrative law, then, determines the level of protection that citizens have, potentially creating tensions in cross-border setting.[173] For example: appealability may differ under national administrative law. That affects decisions such as appointing trusted flaggers, out-of-court dispute settlement bodies, bilateral commitments with providers of intermediary services; all of which are decisions that shape the freedom of expression on the internet. If one member state has strict time limits regarding administrative procedures, but an internet user is waiting for an enforcement decision against a provider of intermediary services in the jurisdiction in which the intermediary service provider is located, it may lead to a situation in which enforcement decisions are not appealable because delays are caused across jurisdictions. This diminishes the protection and right to judicial review of EU internet users.

4. Different enforcement strategies leading to different applications of the DSA

Member States can select existing authorities, found new authorities, or select multiple authorities as DSC; currently, they overwhelmingly opt for appointing existing authorities.[174] The nature of the institution hosting the DSC affects the

[170] Rademacher & Marsch in Hofmann and Raue (n 18) 986–988. Rademacher and Marsch propose a horizontal preliminary ruling procedure under EU secondary law to overcome possible boundaries arising from the foreign review of national law.

[171] Van Cleynenbreugel (n 17).

[172] ibid 9–10; Herwig CH Hofmann, 'Multi-Jurisdictional Composite Procedures - The Backbone to the EU's Single Regulatory Space' (2019) <https://www.ssrn.com/abstract=3399042> accessed 24 October 2024.

[173] See in a data protection context: Gloria González Fuster and others, 'The Right to Lodge a Data Protection Complaint: OK, but Then What?' (AccessNow 2022).

[174] Pieter Van Cleynenbreugel and Pietro Mattioli, 'Digital Services Coordinators and Other Competent Authorities in the Digital Services Act: Streamlined Enforcement Coordination Lost?' [2023] European Law Blog <https://europeanlawblog.pubpub.org/pub/digital-services-coordinators-and-other-competent-authorities-in-the-digital-services-act-streamlined-enforcement-coordination-lost> accessed 19 August 2024. Van Cleynenbreugel and Mattioli suggest that rec. 109's phrasing that Member States "should be able to entrust more than one competent authority, with specific supervisory or enforcement tasks and competences concerning the application of this Regulation" should be interpreted as favouring conferring powers to existing authorities competent in related fields.

application of the DSA in three ways: (i) differences based on available resources, (ii) differences based on enforcement strategies and priorities; (iii) differences based on enforcement traditions of the national authority selected as DSC. Different enforcement strategies indirectly shape the freedom of expression of internet users: they determine the degree to which intermediary service providers can self-regulate, how strictly the DSA is enforced, but also how internet users are protected through effective remedies for potential infringements.[175]

The first instance that causes different levels of enforcement is related to resources and available expertise. Article 50(1) requires that DSCs have the necessary technical, financial, and human resources to supervise all providers of intermediary services in their competence,[176] solidifying sincere cooperation in DSA enforcement between member states.[177] Some DSCs will therefore have to be larger than others, due to the number of providers of intermediary services established in their territory. However, differing levels of resources, particularly under-resourcing, can impact DSA enforcement. Orlando-Salling proposes that the financial resources of DSCs in 'peripheral'[178] member states, against a background of fiscal and debt crises, could strain the enforcement of the DSA in those member states.[179] Orlando-Salling rightly notes that this creates a disbalance in the degree to which the DSA is shaped by 'core' member states, where citizens of 'peripheral' member states could be less protected.[180] As a result DSCs with limited resources may not have the full expertise required to deal with complicated matters at scale. Differences in resources in other sectors with a decentralised enforcement structure, such as competition law and data protection, are found to lead to fragmented enforcement.[181] Being under-resourced can cause (a) case-overload and (b) failure to investigate potentially new infringements, ultimately endangering the protection of fundamental rights – such as the right to freedom of expression in the case of the DSA. For example, in a data

[175] Charter of Fundamental Rights of the European Union, art. 47; Court of Justice of the European Union, 22 December, *DEB v Germany*, C-279/09, para 33. See also De Gregorio and Demková (n 128).

[176] Regulation 2016/679, art. 52(4).

[177] Treaty on the European Union, art. 4(3). It ensures the smooth operating of cooperative mechanisms such as article 57 and 58, but also crucial processes such as the sharing of information through the system developed by the Commission under article 85.

[178] referring to John Reynolds, 'The Political Economy of States of Emergency Symposium: Third World Approaches to International Law (TWAIL) Conference: Capitalism and the Common Good' (2012) 14 Oregon Review of International Law 85.

[179] Orlando-Salling (n 100) 9–10.

[180] Orlando-Salling (n 100); building on Damjan Kukovec, 'Law and the Periphery' (2015) 21 European Law Journal 406.

[181] Jacques Steenbergen, 'Challenges to Enforcers', in Catalin Rusu and others (eds), *New Directions in Competition Law Enforcement*, vol 4 (Wolf Legal Publishers 2020) 1–19; 'Overview on Resources Made Available by Member States to Data Protection Authorities and on Enforcement Actions by the Data Protection Authorities' (European Data Protection Board).

protection setting, the EU Agency for Fundamental Rights notes that insufficient resources[182] has caused DPAs to prioritise certain enforcement tasks causing an inability to take action against potential harms to data subjects.[183] Strowel and Somaini theorise that the Irish DSC will be strained for resources due to the large number of providers of intermediary services established in that Member State.[184] Similarly, underfunding and understaffing is also identified as a barrier in consumer law enforcement.[185] Differences in resources could ultimately affect DSA enforcement as well, and thus how freedom of expression of EU internet users is protected by DSCs.

The second instance that can affect DSA enforcement is varying enforcement strategies. Decentralised authorities can serve national interests in enforcing EU law.[186] Gentile and Lynskey describe how different enforcement strategies led to a fragmented application of the GDPR.[187] Enforcement tasks are interpreted differently across competent authorities: some authorities are more aggressive in enforcement, others are more idle.[188] National interests can play a role in this regard: being a lenient authority can be a way to attract such companies, A-G Bobek observes in *Facebook Belgium*.[189] This may also apply in a DSA context: DSCs may choose to be more lenient in investigating and sanctioning intermediary service providers to maintain their establishment in that jurisdiction. Lenient enforcement can affect internet user's rights, *de facto* sacrificing protection of freedom of expression online for economic gain. These practices can inspire forum-shopping, meaning that

[182] As required in Regulation 2016/679, art. 52(4).

[183] 'GDPR in Practice: Experiences of Data Protection Authorities' (European Union Agency for Fundamental Rights 2024).

[184] Alain Strowel and Laura Somaini, 'Towards a Robust Framework for Algorithmic Transparency to Tackle the Dissemination of Illegal and Harmful Content on Online Platforms' (2021) <https://cdn.uclouvain.be/groups/cms-editors-crides/droit-intellectuel/CRIDES_WP_2_2021_Alain%20Strowel%20and%20Laura%20Somaini.pdf> accessed 17 October 2024; see also Alain Strowel and Jean De Meyere, 'The Digital Services Act: Transparency as an Efficient Tool to Curb the Spread of Disinformation on Online Platforms?' (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 66. Strowel and Somaini's critique is predominantly based on 'What the DPC-Meta Decision Tells Us about the EU GDPR Dispute Resolution Mechanism' <https://iapp.org/news/a/what-the-dpc-meta-decision-tells-us-about-the-gdprs-dispute-resolution-mechanism> accessed 17 October 2024.

[185] Gert Straetmans and Jasper Vereecken, 'Towards a New Balance Between Private and Public Enforcement in EU Consumer Law' [2024] European Review of Private Law 41, 63.

[186] See on this, for example Court of Justice of the European Union, 9 March 2010, *Commission v Germany*, C-518/07, para 35.

[187] Gentile and Lynskey (n 166) 820–821.

[188] González Fuster and others (n 173).

[189] Court of Justice of the European Union, Opinion of AG Bobek of 13 January 2021, *Facebook v Gegevensbeschermingsautoriteit*, C-645/19, para 124.

intermediary service providers determine their establishment based on that stance the DSC in that jurisdiction takes toward enforcement.[190]

A final factor contributing to different enforcement strategies are the enforcement traditions on which appointed DSCs build. While some DSCs, building on an institution concerned with media law, can rely on expertise on matters concerning freedom of expression, other DSCs building on consumer law or traffic authorities, have less experience in this regard. At the very least, these authorities may have different priorities based on the field of law in which they are rooted.[191] Conversely, entities rooted in a media law background may not have enforcement experience to fully exhaust the investigative and sanctioning powers outlined in II.3.c and II.3.d, since their respective field traditionally does not involve dawn raids and hefty fines as competition law or consumer law do.[192] This, again, requires expertise-building on the level of the DSC, which may lead to different enforcements of the DSA.

5. Interim Conclusion

Sections II and III have described the role of the DSC in relation to the internet user's freedom of expression. On the one hand, the DSC shapes the freedom of expression through powers that affect the content moderation process; on the other hand, it protects the freedom of expression of internet users through sanctioning intermediary service providers and enforcing the DSA. In that regard, it is likely that the DSA is enforced differently across member states, because of different national landscapes, interests and administrative procedures. While this is not problematic in itself, the protection of freedom of expression under the DSA can be a concern when its enforcement leads to extraterritorial enforcement of inapplicable laws, or when enforcement strategies and practices harm the protection of freedom of expression of internet users. This supports the assertion in the introduction: there is a seeming disconnect between the profound impact that DSCs can have on freedom of expression of internet users and the safeguards surrounding their role. This is addressed in the next section.

---

[190] Strowel and De Meyere (n 184) 79; Tobias Mast, 'Platform Law as EU Law' (2024) 73 GRUR International 607, 612; Frosio and Geiger (n 89) 44–46. Of course, providers of intermediary services rely heavily on infrastructure, sectoral regulation and tax benefits of member states, and therefore DSA enforcement will not always be a determining factor in determining the main place of establishment. See Gentile and Lynskey (n 166) 822.

[191] Orlando-Salling (n 100) 14.

[192] Kristina Irion, 'The Independence of Media Regulatory Authorities in Europe' (European Audiovisual Observatory 2019) 120. Although most media authorities have a power of inspection, that power can be objected to by the subjected party. Similarly, although media authorities have the capacity to fine, those fines are incomparable to the potential 6% of annual worldwide turnover under art. 52(3) DSA.

IV. Suggestions for better safeguarding freedom of expression in DSA enforcement

This Section describes some avenues for better safeguarding freedom of expression through the application of the DSA. These are by no means exhaustive, and some of the risks identified above are inherent to the chosen enforcement structure in the DSA, and cannot be altered without significant DSA amendments. However, two avenues fitting the existing framework of DSA enforcement may mitigate the risks to freedom of expression outlined above: centralisation in cases of common concern – more so than currently foreseen in Article 56 DSA – to ensure better protection of freedom of expression in such cases; and guidance by the Board in standardising DSA enforcement, on the interpretation of illegal content, and better streamlining exchanges on national law, for more coherent DSA enforcement with regards to freedom of expression.

1. Further centralise DSA enforcement

Firstly, in complex cases with significant fundamental rights impact beyond VLOPSEs, the Commission and/or the Board should be involved in enforcement. DSA enforcement is vertically centralised for VLOPSEs, but some platforms with less than 45 million active monthly users cause similar concerns despite the not meeting the 'simple' threshold of monthly active users.[193] For example, Section II.1 mentioned the relevance of non-VLOPSE platforms such as Telegram or Twitch. Instances in which non-VLOPSEs can affect European elections, involve widespread censorship allegations, or cause significant harm to European citizens across borders, could warrant enforcement beyond the competence of the DSC of establishment. In the context of data protection, which lacks vertical centralisation compared to the DSA, Bastos and Palka suggest that enforcement must be centralised with the Commission or the EDPB in "cases of common European concern"[194] to overcome enforcement deficits and protect constitutional rights to data protection.[195] Despite VLOPSEs already accounting for centralisation – to a degree – in cases of common European

[193] Regulation 2022/2065, art. 33(1). The number of monthly active users has been further specified in the Commission's Guidance on the requirement to publish user numbers: https://digital-strategy.ec.europa.eu/en/library/dsa-guidance-requirement-publish-user-numbers; Under Article 33(2) and (3) the EU legislator expresses the desire for the threshold for VLOPSEs to correspond to roughly 10% of the EU population.

[194] In cases of common European Concern they refer to cases that involve multiple jurisdictions and have a significant potential impact on fundamental rights across the EU.

[195] Filipe Brito Bastos and Przemysław Pałka, 'Is Centralised General Data Protection Regulation Enforcement a Constitutional Necessity?' (2023) 19 European Constitutional Law Review 487.

concern,[196] further centralisation in sensitive cases avoids complications arising from enforcement at the national level described in Section III. Centralising enforcement in sensitive non-VLOPSE cases can ensure that freedom of expression is not harmed by inactive or politically motivated DSCs, or that the application of national law in such cases would create interferences with freedom of expression across the EU. This can be achieved in two ways: vertical centralisation with the Commission, and with the Board.[197]

Attributing enforcement competence beyond VLOPSEs to the Commission in cases of common European concern requires amendment to the DSA. Centralisation is enabled to some extent by the procedure laid down in article 59 on referral to the Commission, but that article only foresees the Commission reviewing enforcement actions taken by national authorities, not actually deciding on enforcement itself. This means that the risk of regulatory capture or administrative paralysis in sensitive cases is still located at the national level. The Board could make a recommendation of whether online platforms have a particular sensitivity toward fundamental rights in the EU that invokes the competence of the Commission, much akin the framework of article 58(2). Complaints by service recipients across the EU are incorporated in assuming this competence, and maintains proximity of the complainant to the procedure that will ultimately affect their freedom of expression, while simultaneously aligning with the wish of the Parliament for more involvement of the Commission in the DSA's enforcement.[198] This ensures that the DSA's application is not nationalised, and 'difficult' platforms do not necessarily need to be supervised by a single DSC; essentially, it sacrifices a degree of member state competence for more efficient enforcement. Of course, there are downsides to this approach: the Commission is not officially attributed competence in these manners, and its competence would require member states ceding additional enforcement powers. This approach may also contravene the principle of mutual trust, because it implies ineffectiveness of national enforcement in particularly sensitive cases.[199] Additionally, the critique of Section II on the independence of the Commission as an enforcement actor in the DSA framework still stands; this suggestion would therefore require a DSA amendment or member states to explicitly cede additional

[196] Regulation 2022/2065, rec. 76.

[197] Of course, the proposed independent European Digital Services Coordinator (see fn. 20) would be a third avenue, but this suggestion departs even further from the existing framework, and is therefore beyond the scope of this contribution.

[198] European Parliament Resolution of 20 October 2020 on the Digital Services Act and fundamental rights issued posed (2020/2022), *OJ C 404*, 6.10.2021, p. 53-62, para 18, 30, 40. See also Rademacher & Marsch Hofmann and Raue (n 18) 985.

[199] Court of Justice of the European Union, 18 December 2014, Opinion 2/13, para 168; Treaty on the European Union, art. 2.

enforcement competence to the Commission. Enforcement in such instances could also be advised by the Board under article 63(1)(d), further solidifying Europe-wide support for the Commission's enforcement action.

Centralising enforcement with the Board could be achieved by allowing the Board to issue binding recommendations on initiating joint investigations. Currently, the Board may recommend joint investigations to the member state of establishment;[200] if the member state of establishment disagrees with the requested enforcement action, they communicate so in their preliminary position.[201] If that is deemed insufficient, the Board may refer the matter to the Commission for dispute resolution under article 59, which would ultimately require the DSC of establishment to report back to the Commission, having taken 'the utmost account' of the Commission's view.[202] Allowing the Board to make binding requests for joint investigations, altering article 60, ensures enforcement in cases of common European concern, even if the DSC of establishment is unwilling to enforce the DSA, without the complexity of an infringement procedure *ex* article 258 TFEU. This would of course affect the procedural autonomy of member states; to somewhat justify that, the DSA would need to be amended, and requests should be justified by a majority or qualified majority of DSCs.

2. Further guide DSA enforcement

A second avenue that should be pursued is better use of guidelines by the Commission and the Board to guide DSA enforcement in alignment with freedom of expression. This is already explicitly established throughout the DSA, for example in the context of the selection of trusted flaggers,[203] and for the Board on the development of European guidelines.[204] Some aspects in the DSA require further guidance by the Commission and the Board. Practical guidelines on ensuring the independence of DSCs from their hosting authority are currently missing; as illustrated in Section II.3.a, the 'thin' independence requirement can blur competences and create questions on the nature of DSA enforcement. Similarly, the Board must coordinate harmonious application of the DSA – the Commission may, through its seat on the Board, nudge the Board in that direction. Guidelines should be issued on interpreting illegal content in article 3(h); these guidelines should

[200] Regulation 2022/2065, art. 60(1)(b).
[201] Regulation 2022/2065, art. 60(2). This is also subject to a requirement of good faith cooperation under art. 60(4).
[202] Regulation 2022/2065, art. 60(3)(b); art. 59(3).
[203] Regulation 2022/2065, art. 22(8).
[204] Regulation 2022/2065, art. 61(3)(e).

address (a) the connecting factor of a set of laws to a particular case;[205] and (b) how the safeguard of compliance with EU law for national law can best be interpreted by actors relying on the definition of article 3(h).

Further, the interpretation of national laws may, in cross-border cases, be dependent on foreign DSCs. In cases of unclarity of the substance of those laws, there are options for sharing information and mutual assistance between DSCs. The *lex ferenda* proposed by Rademacher and Marsch for "the introduction of an explicit horizontal preliminary ruling procedure under EU secondary law" streamlines this even further.[206] If a national court of the competent DSC, or the competent DSC itself, has issues interpreting aspects of a foreign law, they may refer those issues back to a court or the DSC in the requesting member state. This may prevent issues of extraterritorially or wrongly applying national laws identified *inter alia* under III.2, and has the added benefit of allowing the complainant to be in closer proximity, and perhaps even appeal, decisions taken by the court of DSC in their member state in that preliminary reference procedure. The downside of this proposed solution is that the risks of composite procedures, such as delay and administrative paralysis, compound when the composite procedure requires authorities to refer issues back to member states that had originally forwarded the complaint; as such, the expedience of the procedure is an important factor to take into account before relying on Rademacher and Marsch's proposed horizontal procedure.

V. Conclusion

This article has analysed the role of the Digital Services Coordinator under the DSA in light of the freedom of expression of internet users. It highlights how DSCs, on the one hand, shape content moderation practices, and on the other hand how the DSC's role in the decentralised enforcement structure of the DSA affects the protection of freedom of expression of users from platform's content moderation practices. Both of these dimensions ultimately affect the right to freedom of expression of internet users. Safeguards around the protection of that right are sparse in the DSA: as an instrument, it mostly emphasises the responsibility that intermediary service providers have with regards to respecting freedom of expression, but does not acknowledge the risks that enforcement of the DSA at the national level can pose to the freedom of expression of internet users, either by varying speech traditions, or by its decentralised enforcement framework. DSCs only have responsibility toward

[205] Mauritz (n 149).

[206] Rademacher and Marsch in Hofmann and Raue (n 18) 988.

freedom of expression in generally phrased safeguards, without a sufficient risk mitigation mechanism in place to address manifesting harms. This contribution proposes that these risks can be mitigated by a more uniform understanding of 'illegal content' facilitated by horizontal preliminary proceedings, a possible enforcement competence for the Commission or the Board beyond VLOPSEs, and exhaustive use of guidance by the Board, especially on the requirement of independence for DSCs.

DSCs have a set of instruments to affect content moderation, and regulate platforms that ultimately determine user's freedom of expression online. How these instruments are used is therefore key in creating a digital space in which freedom of expression is protected. The DSA talks the talk when it comes to protection of freedom of expression. If the DSA is to really protect freedom of expression online, its enforcement should walk the walk.

Chapter VII: Jawboning Content Moderation from a European Perspective

Published as: Jacob van de Kerkhof, 'Jawboning Content Moderation from a European Perspective' in Charlotte van Oirsouw and others (eds), European Yearbook of Constitutional Law: Constitutional law in the Digital Era, vol 5 (1st edn, TMC Asser Press 2024), 61-85.[1]

To provide context that reflects developments since finalising this work in 2023, this chapter finishes with section 'Afterthoughts', which was not part of the original contribution.

Abstract: *State actors increasingly rely on internet intermediaries to combat unwanted speech types such as hate speech and disinformation. They are regarded as the only entities capable of effectively moderating speech on their respective social media platforms. Despite being private entities, they thereby increasingly are given state-like responsibilities in terms of setting norms for speech on the internet through forms of self- and co-regulation and in terms of enforcing those norms. However, when regulating speech, they are not bound to the same constitutional norms as states are. Their position as an intermediary thus creates a constitutional tension: by informally persuading or pressurizing intermediaries to moderate certain types of speech, states can de facto limit the freedom of speech without adhering to rule of law principles. The figure of informal pressure on private entities to pursue a policy goal that lacks a clear legal basis has been termed 'jawboning' in literature, but the phenomenon is relatively unexplored in the European context. This chapter there- fore aims to explore the constitutional tension that jawboning brings about, taking the European Code of Conduct on Countering Illegal Hate Speech Online as the main example.*

[1] The author is very grateful to prof. Janneke Gerards for her insights and feedback in writing this contribution.

1. Introduction

The internet has provided modern societies with unprecedented access to information. As Benkler has aptly described it, this is the result of three particular characteristics of the internet, which (i) decentralises the production of information; (ii) offers access to information at an extremely low price, placing the means to information in many hands; and (iii) gives an open and global character to communication flows.[2] Clearly, these characteristics make the internet a great asset to the exercise of rights and freedoms such as the freedom of speech and the right to receive information. However, information consumption on the internet is not without risk. Information is largely hosted and spread by private actors on the internet: internet intermediaries.[3] Intermediaries do so with the help of algorithms to sort and present the information to internet users. Algorithms are notoriously opaque and are often said to form a black box.[4] Due to the way algorithms work, they primarily show us information that we 'want' to see.[5] This keeps users 'hooked' to the platform, which is attractive from a commercial perspective, especially as it means that users can be reached by advertisements or commercial content. At the same time, it is well established that the dynamics of algorithm-driven platforms have been shown to lead to the spread of sensationalist and harmful content.[6] Potentially harmful content such as hate speech, disinformation and terrorist propaganda is more likely to be amplified than less harmful content on social media platforms; for example, a study by New York University found that far-right content was more engaging during the 2020 election.[7] In 2014, the World Economic Forum recognised the spread of misinformation online as one of the top 10 trends worldwide, and for good reason.[8] Misinformation online has played a role in the 2014

[2] Yochai Benkler, *The Wealth of Networks: How Social Production Transforms Markets and Freedom Contract : Freedom in the Commons* (Yale University Press 2006).

[3] 'Internet Intermediaries', or in short 'intermediaries' shall be used throughout this paper to predominantly refer to 'very large social media platforms' in the meaning of Article 33 of Regulation 2022/ 2065. The term intermediaries is preferred in this context because it better signifies the role these entities take in the speech governance paradigm.

[4] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (First Harvard University Press paperback edition, Harvard University Press 2016) 9 et seq.

[5] Paddy Leerssen, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (2020) 11 European Journal of Law and Technology 51; Paddy Leerssen, 'An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation' (2023) 48 Computer Law & Security Review 105790.

[6] Siva Vaidhyanathan, *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy* (1st edn, Oxford University Press 2021); Huszár F, Ktena SI, O'Brien C, Bellia L, Schlaikjer A and Hardt M, 'Algorithmic amplification of politics on Twitter' (2022) 119 PNAS e2025334119.

[7] Cybersecurity for Democracy, 'Far Right News Sources on Facebook More Engaging' (*Medium*, 4 March 2021) <https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90> accessed 12 April 2025.

[8] 'Outlook on the Global Agenda 2014' (World Economic Forum 2014).

Crimea annexation, the 2016 US Presidential Elections and the 2020 COVID pandemic. Similarly, online hate speech has been identified by UN Special Rapporteur de Varennes as a real risk for minorities, potentially leading to discrimination, abuse, and in the extreme case of the Rohingya, ethnic cleansing.[9]

In response to these problematic aspects of online speech, the European Commission has created a number of co-regulatory instruments, such as the Code of Conduct on Countering Illegal Hate Speech Online and the EU Code of Practice on Disinformation.[10] These co-regulatory instruments are the result of cooperation between the European Union ('EU') institutions and internet intermediaries. This is a deviation from the traditional dyadic approach to the regulation of speech, which entails that the state determines the outer limits of the freedom of expression but, at the same time, is prevented from limiting freedom beyond its constitutional powers. This dyadic approach is closely related to the core of the rule of law, which is that the power imbalance between citizen and state is tempered by the fact that the state needs to adhere to its own rules for regulation.[11] Compared to this, authors such as Balkin have argued that co-regulatory instruments represent a triadic system of speech governance.[12] State authorities lack enforcing power in the online domain because that domain is largely in private hands. In addition, they have a lack of resources and lack of technical knowledge. Due to these regulatory issues, internet intermediaries are *de facto* governing the digital realm, using instruments such as their terms of service agreements.[13] Importantly, in doing so, internet intermediaries are not bound by the same constitutional confines as the state. Hence, even though social media spaces are private property and do not have any constitutional accountability, their owners fulfil a semi-public function and play a crucial part in

[9] 'Public Comment by UN Special Rapporteur on Minority Issues Relating to Cases on Hate Speech and Minorities' (*OHCHR*) <https://www.ohchr.org/en/statements-and-speeches/2020/12/public-comment-un-special-rapporteur-minority-issues-relating-cases> accessed 12 April 2025.

[10] Code of Conduct on Countering Illegal Hate Speech Online https://ec.europa.eu/newsroom/just/ document.cfm?doc_id=42985 EU Code of Practice on Disinformation 2018, later expanded on by The Strengthened Code of Practice Disinformation 2022, see also Joint Communication to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions Tackling COVID-19 Disinformation–Getting the Facts Right, JOIN(2020) 8 final, 10 June 2020.

[11] Martin Krygier, 'Tempering Power' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017); Sumit Bisarya and W Elliot Bulmer, 'Rule of Law, Democracy and Human Rights: The Paramountcy of Moderation' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017) 128.

[12] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011; Robert Gorwa, 'The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content' (2019) 8 Internet Policy Review 1.

[13] Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598.

the public discourse.[14] This leads to a constitutionally fragile situation, which, according to Balkin may bring a risk of collateral censorship, which means that platforms err on the side of caution and remove legal content to avoid liability, as well as a risk of administrative prior restraint, which means that people do not get to speak until the intermediary gives them permission to do so.[15]

A particular risk arises in the amount of pressure state actors can place on internet intermediaries. By pressuring internet intermediaries to participate in the co-regulation of speech on their platforms, state actors can push internet intermediaries to moderate content in a certain manner.[16] This can be done for many different reasons, ranging from seeking to combat COVID-19 disinformation to trying to silence dissident voices in authoritarian regimes, and in different ways. States can exert such pressure by threatening future regulation in cases of non-compliance, as the EU has done in the case of hate speech, but also by threatening to downright take the website offline in that state, as for example, Turkey has done when it came to blocking political opposition.[17] As a result, the state may *de facto* limit speech through the intermediary. This figure is called 'censorship by proxy.'[18] This allows a state actor to sidestep constitutional safeguards that normally apply in restricting speech, because those same safeguards do not apply to the intermediary. A term coined to describe this censorship-by-proxy through putting informal pressure on intermediaries is 'jawboning'.[19] The term 'jawboning' itself comes from the Biblical story of Samson fighting a thousand men with a jawbone in his hand (Judges 15:15); the analogy being that some great effect can be achieved by something very little. 'Jawboning' is a term used predominantly in American literature in the context of government action pertaining to content moderation. This is explicable in light of the First Amendment to the US Constitution, which is markedly more limitative toward government involvement in speech than Article 10 of the European

[14] Evelyn Douek, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 526.

[15] Balkin, 'Free Speech Is a Triangle' (n 12) 2016–2018.

[16] Daphne Keller, 'Who Do You Sue? State and Platform Hybrid Power over Online Speech' [2019] Hoover Working Group on National Security, Technology and Law.

[17] See for example ProPublica 2021 https://www.propublica.org/article/sheryl-sandberg-and-top- facebook-execs-silenced-an-enemy-of-turkey-to-prevent-a-hit-to-their-business.

[18] Seth F Kreimer, 'Censorship by Proxy: The First Amendment, Internet Intermediaries, and The Problem of The Weaker Link' (2006) 155 University of Pennsylvania Law Review 11. coins the term censorship by proxy. Balkin also uses the term 'collateral censorship' to mean jawboning when collateral censorship is the result of exercise of soft power, see Jack M Balkin, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296, 2298; 2327.; Amilie Pia Heldt, 'Bypassing Classical Concepts of Censorship?' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 56, 62. Heldt uses this same definition: 'collateral censorship is when public authorities force intermediaries to control their users communication.' There is a slight distinction between the two approaches which is discussed below.

[19] Derek Bambauer, 'Against Jawboning' (2015) 100 Minnesota Law Review 51.

Convention of Human Rights is, and thereby may invite some sidestepping. In view of the value of the freedom of speech and the right to information, however, it is considered problematic if a government would evade the nearly absolute protection of the First Amendment by using informal means to effectively 'abridge' the freedom of speech.[20]

In a European context, jawboning has been given less attention.[21] Indeed, European constitutional norms provide more possibilities to regulate freedom of expression, provided that restrictions are sufficiently justified and comply with several constitutional norms.[22] European human rights documents, notably the European Convention on Human Rights (ECHR or Convention) and the EU Charter of Fundamental Rights (CFR or Charter) expressly provide grounds and formulate requirements for limiting the freedom of expression.[23] At first glance, and different from in the US, these possibilities for restriction would also seem to allow for a certain degree of indirect restrictions of the freedom of expression, as long as they meet the requirements set by the relevant provisions. Nevertheless, also here, some specific issues may arise when jawboning is used as a policy instrument, in particular when viewed from a constitutional and rule of law perspective.

This chapter aims to explore the constitutional and rule of law risks of 'jawboning' when used to pursue speech policy aims in the European context. Since authors have used the concept of jawboning in various ways and it has been specifically developed in the United States, it first is useful to offer some further conceptual clarification (Sect. 2). This chapter then outlines the current protection offered by European instruments to the freedom of expression for content carried by internet intermediaries, focusing on criteria for limiting that freedom, as well as the overarching principle of the rule of law (Sect. 3). In light of these criteria, Sect. 4 identifies tensions that arise from 'jawboning' in the European framework of freedom of expression. Finally, Sect. 5 discusses one particular example of jawboning in European speech regulation: the EU's Code of Conduct on Countering

[20] ibid 51.

[21] With the exception of Danielle Keats Citron, 'Extremist Speech, Compelled Conformity, and Censorship Creep' (2018) 93 Notre Dame Law Review 1035; Patrick Leerssen, 'Cut out by the Middle Man: The Free Speech Implications of Social Network Blocking and Banning in the EU' (2015) 6 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 99; Michael Karanicolas, 'Squaring the Circle Between Freedom of Expression and Platform Law' (2020) 20 Pittsburgh Journal of Technology Law & Policy 177.

[22] E.g. Article 5(2) Grundgesetz; Article XI Déclaration des droits de l'homme et du citoyen, see also Stephen Gardbaum, 'The Structure of a Free Speech Right' in Adrienne Stone and Frederick Schauer (eds), *The Oxford Handbook of Freedom of Speech* (Oxford University Press 2021).

[23] Article 10 ECHR(2); Article 52(1) CFR.

Illegal Hate Speech Online.

2. Definitions: What Do We Mean by 'Jawboning'?

Generally, jawboning falls in a wider category of informal regulatory instruments that create tensions from a constitutional standpoint.[24] One definition, offered by Bambauer, is that jawboning is 'a specific type of informal pressure by a government actor on a private entity: one that operates at the limit of, or outside the actor's authority.'[25] This definition also finds expression in various other terms that are used in literature to describe the meaning of this and related phenomena, such as 'moral suasion',[26] 'regulation by raised eyebrow',[27] 'regulation by invisible hand- shake,'[28] or, specifically in speech regulation, 'censorship by proxy'[29] and 'collateral censorship'.[3031] Bambauer has created a taxonomy of jawboning phenomena along a two-dimensional scale: one axis being related to coercion, the other to authority. 'Coercion' is described as a threat or sanction that actors face when refusing to comply with informal pressure. 'Authority' is the legal basis of the informal pressure and relates to the competence of the state actor. The idea behind the taxonomy is that if coercion is stronger, and the possible sanctions more pressing, the government actor requires more authority to apply that coercion. Along this two-dimensional graph of governmental pressure, jawboning exists at the intersection of 'no authority' and (a gliding scale of) 'coercion'. Seen from a rule of law perspective, such coercion and sanctioning without authority is problematic, especially when it possibly affects the

[24] Alberto Alemanno and Anne-Lise Sibony, *Nudge and the Law : A European Perspective* (1st edn, Hart Publishing 2015); Anne van Aaken, 'Constitutional Limits to Nudging: A Proportionality Assessment' in Alexandra Kemmerer and others (eds), *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging* (Hart 2015).

[25] Bambauer (n 19) 56.

[26] JT Romans, 'Moral Suasion as an Instrument of Economic Policy' (1966) 56 The American Economic Review 1220, 1220.

[27] Yochai Benkler, 'A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate' (2011) 46 Harvard Civil Rights Civil Liberties Law Review 311.

[28] Michael Birnhack D and Niva Elkin-Koren, 'The Invisible Handshake: The Reemergence of the State in the Digital Environment' (2003) 8 Virginia Journal of Law & Technology 1.

[29] Kreimer (n 18).

[30] Balkin describes 'collateral censorship' as follows: 'The first feature is collateral censorship, in which the state regulates party A in order to control speaker B. The digital age enables a vast number of people to communicate widely across the country and around the world. Because there are so many speakers, who are often anonymous, difficult to co-opt, or otherwise beyond the government's effective control, the state aims at Internet intermediaries and other owners of digital infrastructure–threatening liability to induce them to block, limit, or censor speech by other parties'. See Balkin, 'Old-School/New-School Speech Regulation' (n 18).

[31] Outside of speech regulation, it is also used in the context of inflation management, labour affairs and climate policy. Indeed, as early as the seventies, the term jawboning was used to describe a 'behind the scenes' presidential intervention to achieve a policy goal. Verkuil, in his article Jawboning Administrative Agencies: Ex Parte Contacts by the White House, notes that jawboning has been a part of the presidential repertoire since Kennedy 'sought to restrain prices and wages in the steel industry. William Safire, in his Safire's Political Dictionary of 1978, equates it to moral suasion, the term used to describe jawboning during the Carter Administration.

constitutional rights of citizens.

At the bottom of the coercion scale, one could find reputational sanctions–a relatively innocent sanction, for which not a lot of authority is required. For example, when Wikileaks disclosed large amounts of diplomatic information, the Obama administration pressured infrastructure providers such as Amazon (Wikileaks' webhost) and Paypal (Wikileaks' donation mechanism) to stop supporting Wikileaks, because their support would be unpatriotic.[32] No legal basis is required for this pressure, but it can still raise a constitutional concern. An intermediary might suffer reputational damage for non-compliance, but compliance might infringe on their rights. Moral suasion is another term in literature coined to describe this phenomenon. Romans identifies the following four elements for moral suasion: (i) coercion of a (ii) private actor by a (iii) governmental actor to take action that is not (iv) legally enforceable as of yet or part of an existing body of law.[33] In the purest sense, moral suasion is an appeal to altruistic behaviour, without coercive elements. Such 'pure' moral suasion, as Romans coins it, would be effective in bouts of wartime, for example, alluding to patriotism.[34] In speech governance, this is visible in the EU relying on internet intermediaries to block access to Russian news outlets to limit the spread of Russian war propaganda.[35] If an intermediary fails to meet the appeal to altruistic or patriotic behaviour, the sanction is not necessarily future regulation, but there is certainly a heavy reputational sanction. This is similar to 'regulation by raised eyebrow', mentioned above as a species of jawboning. This figure is used to describe the Federal Communications Commission ('FCC') practice of informally pressuring broadcasters, by coercing them via fear and bureaucratic hassle.[36] Although these are relatively innocent examples of jawboning, they can still have significant effects for the intermediary under pressure, raising the question whether a more formal approach is preferable.

---

[32] Derek Bambauer, 'Orwell's Armchair' (2012) 79 University of Chicago Law Review 863, 863; Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 University of California, Davis Law Review 1149.

[33] Romans (n 26) 1222.

[34] ibid 1221. Romans further notes that the notion of altruism in this definition is not inherent to moral suasion: 'the basic difference between moral suasion and direct suasion via the 'rule of law' is not the threat of punishment or promise of reward. Rather, it is that punishment does not
automatically and uniformly fall upon the noncomplier to moral suasion as it does upon the violator of statute law.'

[35] Foo Yun Chee, 'EU's Breton Wants Google, YouTube to Ban War Propaganda Accounts' *Reuters* (27 February 2022) <https://www.reuters.com/technology/eus-breton-hold-video-call-with-google-youtube-ceos-ukraine-fake-news-2022-02-27/> accessed 28 November 2022.

[36] Erwin G Krasnow and M Wayne Milstead, 'FCC Regulation and Other Oxymorons Revisited' (1999) 7 Media Law & Policy 7, 7.

At the other end of the coercion spectrum is a threat of, or even punishment by, state action for non-compliance, for example through fines or regulation. Bambauer mentions various instances of the Obama administration pressuring intermediaries by threatening increasing copyright legislation if the intermediaries failed to act in a way supporting the administration's policy.[37] For the EU, Citron has argued that a large body of EU co-regulatory documents is the result of the EU threatening platforms by limiting their liability immunity for illegal content.[38] There have been many calls from European officials on social media platforms to combat hate speech and terrorist propaganda online, resulting for example in the Code of Conduct on Countering Illegal Hate Speech Online ('Code of Conduct'). Subsequent threats of adopting actual legislation were ousted to ensure better compliance with the Code of Conduct.[39] Informal coercion can be made more substantial by threatening that such legislation may contain the rule that fines can be imposed for hosting illegal content, such as currently already included in the Digital Services Act.[40] Of course, it is possible that the interests of the internet intermediary and the state align in cases of content moderation. Many platforms want to take a stand against harmful content, even if it is only to secure their advertising sales.[41] However, co-optation between a state and an internet intermediary may create an unholy alliance that may hamper speech on the semi-public forum that social media platforms have become. Balkin warns that such co-optation may lead to increasing governing, surveilling and regulation by internet intermediaries.[42] This in turn may make internet intermediaries an even more likely target of jawboning.

In content moderation, this is especially relevant when actions informed by state pressure affect the right to freedom of expression. The question is whether state involvement in those policies leads to accountability for the fundamental rights risks following from jawboning, despite the fact that the intermediary is the main entity limiting speech. To that end, the next section outlines the legal framework around limiting speech through intermediaries in the European legal context, in which it

[37] Bambauer (n 32) 895–896.

[38] Citron (n 21) 1043.

[39] Liat Clark, 'Facebook and Twitter Must Tackle Hate Speech or Face New Laws' *Wired UK* <https://www.wired.co.uk/article/us-tech-giants-must-tackle-hate-speech-or-face-legal-action> accessed 1 November 2022.

[40] Regulation 2022/2065, art. 74.

[41] Tarleton Gillespie, *Custodians of the Internet* (1st edn, Yale University Press 2018) 45–75.

[42] Balkin, 'Free Speech Is a Triangle' (n 12) 2019–2021.: 'The result is a burgeoning dialectic of governing power and public-private cooperation. Private-infrastructure companies develop ever greater governing capacities. Nation-states attempt to co-opt these capacities through coercion or threats of regulation. This, in turn causes increased development of governing, surveilling, and regulatory capacities. And this, in turn makes private-infrastructure owners even more tempting targets for government pressure– because private companies can no longer pretend that they cannot actually do what governments want them to do.'

describes the authority and sanction aspects of limiting speech.

3. Freedom of Expression, Internet Intermediaries and the Rule of Law

Now that the concept of jawboning has been clarified, this section aims to explain the scope of protection of freedom of expression in the ECHR, as well as possible limitations to that protection, with particular attention to internet intermediaries. Section 3.1 focusses on Article 10 ECHR. Section 3.2 pays attention to EU regulation and case law on internet intermediaries that affect freedom of expression, for example, monitoring obligations. Section 3.3 addresses the rule of law as an overarching principle.

3.1 Freedom of Expression and the European Court of Human Rights

Article 10(1) ECHR provides the right to freedom of expression: 'Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers […]'. As the European Court of Human Rights ('ECtHR') has held, the provision has a wide reach and even covers expressions that may 'offend, shock or disturb the State or any sector of the population.'[43] This also applies to internet: expressions on the internet include the use of platform affordances such as 'like' buttons as well as platforms' content moderation policy.[44]

An interference with the right to freedom of expression is only compatible with Article 10 ECHR if it is provided for by law and if it is necessary for a democratic society to pursue one of the legitimate aims laid down in para 2 of the article.[45] 'Provision by law' is a rule of law component: the exercise of state power in the freedom of expression must be accessible and foreseeable.[46] The principle of 'necessity in a democratic society' is a requirement of proportionality, which implies that the interference with the right to freedom of expression can only be regarded as justified if the right is outweighed by a pressing social need, and the interference is a suitable and necessary means to achieve this end.[47]

---

[43] European Court of Human Rights, *Handyside v the United Kingdom,* judgement of 7 December 1976, 5493/72, para 49

[44] European Court of Human Rights, *Melike v Turkey,* judgement of 15 September 2021, 35786/19, European Court of Human Rights, *Delfi v Estonia,* judgement of 16 June 2015, 64569/09; see on affordances Taina Bucher and Anne Helmond, 'The Affordances of Social Media Platforms' in Jean Burgess, Alice Marwick and Thomas Poell, *The SAGE Handbook of Social Media* (SAGE Publications Ltd 2018).

[45] Article 10(2) ECHR; EM Barendt, *Freedom of Speech* (2nd ed, Oxford University Press 2007).

[46] European Court of Human Rights, *The Sunday Times v the United Kingdom,* judgement of 26 April 1979, 6538/74, para 49.

[47] European Court of Human Rights, *Animal Defenders International v the United Kingdom,* judgement of 22 April 2013, 48876/08, para 100.

Then applied to content moderation by internet intermediaries, this framework presents a challenge: the ECHR and the CFR traditionally govern the relation- ship between citizen and state, not dissimilar to national constitutional provisions. However, as explained in the introduction, the regulatory paradigm for content moderation and speech governance on the internet as laid down above is triadic, as it concerns the user, the internet intermediary and the state.[48] This creates a difficulty, since internet intermediaries are not bound to uphold the freedom of expression under these international treaties.[49] Moreover, under EU law, content carried by an internet intermediary cannot be considered a communication act of that internet intermediary, unless the internet intermediary has knowledge of the content's existence and performs a role exceeding passive, technical and automated.[50] This clearly raises some issues of responsibility as regards the information that is spread through the internet. The ECtHR has tried to solve this by stating that even though an internet intermediary cannot be considered the author of speech it hosts, it has responsibility for that content, especially if it has control over it.[51] Especially larger internet intermediaries should assume responsibility in this respect. This may even extend to particular users of social media platforms whose pages host illegal comments, in particular politicians.[52] Such liability of intermediaries for user-generated content makes sense from a practical standpoint: the users of social media platforms are often anonymous, and tracking them down would prove an unpractical exercise.[53] Further, intermediaries have the potential to govern speech at a larger scale than individual decisions.[54] It is, therefore, an attractive policy option for states to target the internet intermediary, as, in a sense, the ECtHR has also done in its judgements. Conversely, access to internet intermediaries is also recognised as an exercise of the right to freedom of expression. For example, blocking access to an internet intermediary was considered an infringement of Article 10 ECHR in

[48] Balkin, 'Free Speech Is a Triangle' (n 12).

[49] Douwe Korff, 'The Rule of Law on the Internet and the Wider Digital World' (Council of Europe 2014) Issue Paper. One could argue that they are bound to respect human rights by the UN Guiding Principles on Business and Human Rights, yet those principles are non-binding.

[50] Article 6 Regulation 2022/2065 on a Single Market for Digital Services, Court of Justice of the European Union, *Sabam v Netlog*, judgement of 16 February 2012, C-360/10.

[51] European Court of Human Rights, *Delfi v Estonia*, judgement of 16 June 2015, 64569/09, paras 144–145, European Court of Human Rights, *MTE v Hungary*, judgement of 2 February 2016, 22947/ 13, paras 62–63, European Court of Human Rights, *Sanchez v France*, judgement of 15 May 2023, 45581/15.

[52] European Court of Human Rights, *Sanchez v France*, judgement of 15 May 2023, 45581/15.

[53] European Court of Human Rights, *Delfi v Estonia*, judgement of 16 June 2015, 64569/09; Teresa Quintel and Carsten Ullrich, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond' in Bilyana Petkova and Tuomas Ojanen, *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020).

[54] Balkin, 'Free Speech in the Algorithmic Society' (n 32).

Yilderim v Turkey and subsequent cases.[55] This doctrine was extended in *Delfi v Estonia*, in which the Court found that issuing injunctions toward an intermediary to monitor content in a comment section could be considered an infringement of the freedom of expression of the user and the intermediary, if not justified under Article 10(2) ECHR.[56] In *Delfi*, such monitoring was justified due to the nature of the comments that Delfi was fined for, containing defamatory content.

Delfi caused a shockwave amongst websites that hosted user-generated comment sections, as they feared that the case would lead to an *ex ante* moderation requirement.[57] The findings in *Delfi* were somewhat nuanced in *MTE v Hungary*, however, in which the Court found that the right to freedom of expression of the website user could outweigh the rights of the people in the comment section as long as the utterances did not amount to hate speech.[58] In *MTE*, the Court acknowledged that a proactive monitoring obligation to avoid liability could have a chilling effect on freedom of expression, which outweighed the interest of the general public in having the comments removed from MTE's website–as opposed to the *Delfi* case.[59] Moreover, in both cases, the Court accepted that having a notice-and-takedown system on the website could balance the right to respect for private life and the right to freedom of expression.[60] Nevertheless, both judgements make clear that in exceptional instances, an internet intermediary may be required to act, for example in cases of defamatory content.[61]

As briefly mentioned above, the Court has found more recently that this liability also extends to users who make their social media presence public and host third-party comments containing hate speech. In *Sanchez v France*, a local politician suffered criminal liability for not promptly deleting comments containing hate speech from his public Facebook wall.[62] The criminal liability was considered an interference with freedom of expression, but the Court regarded it as justified, since the hate speech in

[55] European Court of Human Rights, *Ahmet Yilderim v Turkey*, judgement of 18 December 2012, 3111/10, European Court of Human Rights, *Cengiz et al. v Turkey*, judgement of 1 December 2015, 48226/10 and 14027/11, European Court of Human Rights, *OOO Flavus et al. v Russia*, judgement of 16 November 2020, 12468/15.

[56] European Court of Human Rights, *Delfi v Estonia*, judgement of 16 June 2015, 64569/09. *See also more recently* European Court of Human Rights, *Sanchez v France*, judgement of 15 May 2023, 45581/15, extending that doctrine to social media pages.

[57] Christopher Marsden, 'Internet Liability Law Case Study' in Madeleine De Cock Buning and Linda Senden (eds), *Private Regulation and Enforcement in the EU: Finding the Right Balance from a Citizen's Perspective* (Hart Publishing 2020).

[58] European Court of Human Rights, *MTE v Hungary*, judgement of 2 February 2016, 22947/13.

[59] Ibid., para 86.

[60] Ibid., para 91.

[61] Robert Spano, 'Intermediary Liability for Online User Comments under the European Convention on Human Rights' (2017) 17 Human Rights Law Review 665.

[62] European Court of Human Rights, *Sanchez v France*, judgement of 15 May 2023, 45581/15.

the comments was so manifestly unlawful that the politician should have taken responsibility in his role as intermediary. The Court emphasised the public role Sanchez had as a politician in election time. While keeping a public Facebook profile during election time, he should have a high level of vigilance in overseeing comments on his profile, especially in light of the fact that the Court considers him an 'active' user.[63] The specific circumstances of this case led the Court to find that the French Court was justified in limiting Sanchez's freedom of expression by holding him criminally liable for comments posted on his Facebook page.

These judgements show that in Europe, limiting speech on the internet by an intermediary is considered an interference with the right to freedom of expression, which may be justified under the requirements of Article 10(2) ECHR, for example if the litigious content is manifestly unlawful.

3.2 Freedom of Expression and the European Union

Freedom of expression on a European level is governed by two systems of fundamental rights. Next to the ECHR, which was discussed in the previous section, the EU and its member states also have to comply with the EU Charter of Fundamental Rights when they implement EU law.[64] Following Article 52(3) CFR, the Court of Justice of the European Union ('CJEU') can rely on the ECtHR's interpretations when the rights of the CFR and the ECHR overlap, as is the case for Articles 10 ECHR and the corresponding Article 11 CFR. Nevertheless, the CJEU has handed down a number of judgements of its own concerning internet intermediaries and freedom of expression. In *Eva Glawischnig-Piesczek v Facebook Ireland*, the CJEU held that the e-Commerce Directive (Directive 2000/31) should be interpreted in such a way that courts could issue injunctions ordering internet intermediaries to proactively monitor content *identical* to the litigious defamatory content.[65] Because defamatory content is often available in multiple forms, internet intermediaries can also be ordered to proactively monitor *similar* content.[66] In this case, the CJEU implied that the freedom of expression of internet users must be balanced against the damage suffered by defamatory content. If the right to respect for private life prevails in such cases, this can lead to filtering obligations for internet intermediaries related

[63] Ibid., paras 193 et seq.

[64] Oreste Pollicino, 'Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis' in Andrej Savin and Jan Trzaskowski, *Research Handbook on EU Internet Law* (Edward Elgar Publishing 2014).

[65] Court of Justice of the European Union, *Glawischnig-Piesczek v Facebook Ireland,* judgement of 3 October 2019, C18/18

[66] Ibid., paras 45–46

to *identical* and *similar* content.[67]

In addition to the general provisions of the Charter, the EU legislature has adopted several legislative acts to counter hate speech, misinformation and other problematic types of speech on the internet. Of particular importance in this regard are the e-Commerce Directive and the Digital Services Act, which regulate the activities of internet intermediaries.[68] These acts prohibit states from imposing general monitoring obligations on internet intermediaries, which are obligations to monitor all hosted content or all circumstances indicating illegal or unwanted content. One of the reasons behind this prohibition is the protection of freedom of expression.[69] General monitoring obligations may inadvertently create collateral censorship; catching the proverbial dolphins in the net.[70] Algorithmic moderation can be inaccurate in moderating content in context, running the risk of 'misinterpreting' content and blocking it even though it is legal and compliant with terms of service agreements. General monitoring to remove illegal types of content therefore invariably balances freedom of expression against the potential harms of the content that is monitored for. Similarly, Directive 2019/790 does not allow for general monitoring obligations in relation to expressions that affect intellectual property (IP) rights.[71] In *Sabam v Netlog* and *Sabam v Scarlet Extended*, the CJEU held in this respect that the right to freedom of expression outweighed the interest of the IP-rights holder.[72]

Except for such general prohibitions, EU law allows for some measures to nudge internet intermediaries into taking a more active role in monitoring illegal content online. For example, with regard to the protection of intellectual property, Article 17(4) of Directive 2019/790 creates a liability regime for internet intermediaries who do not apply high industry standards of due diligence in combatting infringing

[67] Daphne Keller, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' (2020) 69 GRUR International 616.

[68] Directive 2000/31/EC, rec. 47, Article 16; Regulation 2022/2065, rec. 30, Article 8.

[69] Martin Senftleben and Christina Angelopoulos, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the e-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' [2020] Amsterdam/Cambridge.

[70] Daphne Keller, 'Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion' (2019) <https://cyberlaw-stanford-edu.proxy.library.uu.nl/files/Dolphins-in-the-Net-AG-Analysis.pdf> accessed 25 November 2021; Felix T Wu, 'Collateral Censorship and the Limits of Intermediary Immunity' (2011) 87 Notre Dame Law Review 293; Giovanni Sartor, 'The Impact of Algorithms for Online Content Filtering or Moderation' (European Parliament 2020).

[71] Directive 2019/790 Article 17(8).

[72] Court of Justice of the European Union, *Sabam v Scarlet*, judgement of 24 November 2011, paras 49–52, Court of Justice of the European Union, *Sabam v Netlog*, judgement of 16 February 2012, C-360/10, paras 47–48.

content, which is a strong nudge toward active monitoring.[73] In these cases, the interest of the IP-rights holder outweighs *inter alia* the freedom of expression. The EU regulator also incentivises an active role for intermediaries in Article 7 Digital Services Act ('DSA'), removing liability under Articles 4–6 DSA for becoming aware of illegal content discovered during good faith-monitoring. As such, it nudges platforms to actively search for illegal content without running the risk of becoming liable when learning about such illegal content.

3.3 Rule of Law as an Overarching Principle

The judgements and approaches described above can be summarised as follows. Some means of effectively governing speech on the internet–such as upload filters and general monitoring practices–can infringe the freedom of expression of internet users when a state incentivises internet intermediaries to employ them without constitutionally justifying their use. Fundamental rights norms thus stand in the way of states achieving their policy aims on the internet. To avoid being held to account before a court because of their regulation, states may try to use other means to govern freedom of speech. As was mentioned above, they may do so for objective reasons. Indeed, they may genuinely want to prevent hate speech or disinformation from being spread by engaging the third player in the internet triad of provision of information: the internet intermediary. It cannot be excluded, however, that some states use forms of (soft) coercion towards intermediaries for less noble reasons.

Moreover, it can be seen that certain rule of law risks arise. It is accepted that in traditional, binary and vertical relationships, the rights of the citizen are protected from the state's exercise of power by constitutional rights and various rules of law. Indeed, preventing and countering the risk of arbitrariness is the very foundation of curtailing government power through constitutionalism.[74] The principle of constitutionalism concerns "the ways the exercise of public power is constituted, made up."[75] A constitution attributes public power to certain institutions, and then sets limits and conditions for the exercise of those powers by the government vis-à-vis citizens.[76] The principle of governing the exercise of government power can be

[73] Senftleben and Angelopoulos (n 69).

[74] Timothy Endicott, 'The Coxford Lecture: Arbitrariness' (2014) 27 Canadian Journal of Law & Jurisprudence 49, 49; Martin Krygier, 'Four Puzzles about the Rule of Law: Why, What, Where? And Who Cares?' (2011) 50 NOMOS: American Society for Political and Legal Philosophy 64, 64.

[75] Krygier (n 11).

[76] Some constitutions claim to describe the nature of a political order rather than found one, especially older ones. However, modern constitutionalism is about 'the establishment and regulation of the powers of the main institutions of government, to

considered the 'rule of law'. Rule of law is defined in this chapter as 'fundamental rights protection, judicial review, the division of powers, as well as a variety of governance requirements–values that are in some form also legally protected by constitutional norms.'[77] One of those values is legality, which relates to foreseeability and accessibility.[78] This entails that norms should be accessible and clear before they can be considered legal–otherwise they may be applied arbitrarily. The interplay between constitutionalism and the rule of law thus allows for both creating government power and protecting citizens from it by forming a framework for the way in which that power is exercised. Krygier has coined the term *tempering* to define that interplay: government power is tempered by constitutionalism and the rule of law, meaning that it is not only created, but also regulated and guided by those principles.[79]

Regarding fundamental rights, constitutionalism and rule of law are two sides of the same coin. The constitution provides governments with the power to regulate, but this is tempered to the extent that the constitution also sets particular requirements for this and offers several constitutional safeguards, such as judicial protection. Only if a measure is provided by law, serves a legitimate aim and is necessary and proportionate to achieving those aims, and only if sufficient protection is offered against arbitrary use of regulatory and executive power, does the government have the competence to implement the measure that violates the fundamental rights.

The salient question is how these constitutional and rule of law principles, and the concomitant protection of fundamental rights, can be maintained in the type of triadic relations that are central to this chapter and, in particular, in the situation that state authorities engage in jawboning. This question is addressed in more depth in Sect.4.

3.4 Intermediate Conclusion

The right to freedom of expression as guaranteed by Article 10 ECHR also covers speech on the internet. The right can be limited, but such limitations need to be provided for by law, serve a legitimate aim and be necessary and proportionate to

specify the relationship and citizen, and to take effect as fundamental law.' See Martin Loughlin, 'The Constitutional Imagination: The Constitutional Imagination' (2015) 78 The Modern Law Review 1.

[77] Maurice Adams, Ernst Hirsch Ballin and Anne Meuwese, 'The Ideal and the Real in the Realm of Constitutionalism and the Rule of Law: An Introduction' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017); Korff (n 49).

[78] Friedrich Hayek, 'Decline of the Rule of Law' [1953] The Freeman 518.

[79] Krygier (n 11) 46.

achieving this aim. The right can be balanced against other fundamental interests, which are listed in Article 10(2) ECHR. The limitation of freedom of expression can take place directly at the level of the speaker but can also concern the intermediary carrying the speech. In several cases, both the CJEU and the ECtHR have decided that internet intermediaries can be required to take action against defamatory or IP-infringing content. Those interests can outweigh the freedom of expression of the speaker and the internet intermediary.

It is understandable that internet intermediaries are often held responsible for taking action against illegal or harmful types of speech; internet users are often anonymous, making it unfeasible to hold them to account. However, overly general approaches to online speech have been shown to be at odds with the prohibition on general monitoring obligations present in the Digital Services Act and the Directive on Copyright in the Digital Single Market, precisely because they may intrude too deeply into the freedom of expression.

4. Jawboning, Content Moderation and the Rule of Law

This section brings the findings of the previous two sections together by explaining the incentives for jawboning in content moderation and highlighting forms of jawboning in content moderation. It concludes with the constitutional concerns attached to the practice of jawboning in content moderation.

4.1 Incentives for Jawboning in Content Moderation

Speech governance requires a balancing act between state authorities pursuing policy goals that serve communal interests and the rights of the individual to freedom of expression and information. When a government seeks to restrict the freedom of expression, it has to comply with the prerequisites to do so outlined in Sect. 3. Those prerequisites are generally strict, and there is a strong presumption in favour of freedom of expression in Europe. Indeed, it has turned out that not all policy goals can be achieved within these fundamental rights safeguards. Oft-discussed examples of this are difficult to discern content types such as defamatory content, some types of hate speech and disinformation.[80] In those cases, it is not always obvious that the policy interest in limiting these 'lawful but awful' content types outweighs the right of the individual to that freedom,[81] or the right of the public to be informed.[82]

[80] Giovanni Pitruzzella and Oreste Pollicino, 'The Constitutional Perspective on Freedom of Expression in the Internet Era' in Andrew Spannaus (tr), *Disinformation and hate speech: a European constitutional perspective* (First edition, Bocconi University Press 2020).

[81] For example, European Court of Human Rights, *Perincek v Switzerland*, judgement of 15 October 2015, 27510/08, para 208.

As seen above, states may turn to internet intermediaries to aid them in governing speech online. This has two reasons: firstly, the intermediary is the only entity with the power to effectively moderate speech on the internet–after all, the social media platform is their property–, and secondly, the internet intermediary is able to block speech acts that a state would be constitutionally precluded from prohibiting. States may order internet intermediaries to remove content, but as explained in Sect.3.2, such orders should not impose a general monitoring obligations on the intermediary; hence, such orders are limited to specific and clearly identifiable pieces of content.[83] By contrast, being a private actor, the internet intermediary is unburdened by constitutional concerns, and only needs to adhere to the terms of service vis-à-vis the user and the formal legislation governing its behaviour, such as the e-Commerce Directive and the Digital Services Act.[84] To a certain degree, states can influence or even coerce internet intermediaries to make use of their possibilities to monitor and govern speech acts for particular purposes. This would allow states to sidestep constitutional constraints on removing 'lawful but awful' content, as is illustrated in the next section.

4.2 What Does Jawboning in Content Moderation Look Like?

Jawboning in content moderation may take place in a number of ways. The consistent factor is the exertion of informal pressure to achieve a policy goal that lacks a clear legal basis. Ordinarily, internet intermediaries set their own rules and standards for operation. Most internet intermediaries thereby engage in some form of self-regulation, setting norms for the use of their platforms, for instance regarding prohibited speech. States can get involved in this process of self-regulation both in formal and informal ways. For instance, as identified by Angelopoulos *et al*., the state can facilitate meetings between stakeholders, it can influence the substance of the measures, for example by instructing companies to take measures or participate in discussions between stakeholders, or it can help in the execution of self-regulation, for example by signalling possible infringements of the private agreement to the

[82] European Court of Human Rights, *Cetin et al. v Turkey*, judgement of 13 February 2003, 40153/ 98, para 64; see also Gerards 2023, p. 84.

[83] Regulation 2022/2065, rec 66, Article 8; see also Court of Justice of the European Union, *Glawischnig-Piesczek v Facebook Ireland*, judgement of 3 October 2019, C-18/18.

[84] Edoardo Celeste, 'Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?' (2019) 33 International Review of Law, Computers & Technology 122, 122. *cf.* Article 14(4) Regulation 2022/2065, stipulating that terms and conditions of large online social media platforms need to be applied with due regard for fundamental rights.

participants in that agreement.[85] In some cases, this results in co-regulation rather than self-regulation.[86] The question is to what degree a state can be involved in regulating before it amounts to actual 'state action' that, as such, has to comply with rule of law and constitutional requirements.[87] An overly narrow concept of 'state action' in this case could diminish the level of fundamental rights protection, and does not sit easily with horizontal and diagonal applications of freedom of expression in the ECtHR's jurisprudence.[88] This aligns with the central argument of this contribution: if informal pressure is exerted by a government to achieve a policy goal for which it lacks formal power, this creates constitutional tension and may violate fundamental rights if there are no checks and balances to ensure that an interference is proportional, necessary and legitimate.

Some forms of this tension are visible in the co- and self-regulatory instruments that are in place in the EU and which affect the freedom of expression. In particular, the EU requests that private entities target harmful and illegal types of content, taking a proactive approach in their regulation and moderation. Examples are the Code of Conduct on Countering Illegal Hate Speech Online, the Communication on Tackling Illegal Content Online, the Recommendation on Measures to Effectively Tackle Illegal Content Online, the Communication on Tackling Disinformation Online, and the Strengthened Code of Practice on Online Disinformation.[89] By means of these instruments, the EU facilitates and takes part in discussions between stakeholders. Most of them further reflect policy wishes regarding harmful content. Platforms usually accept the EU's intervention in their self-regulatory efforts for multiple reasons. Firstly, it allows them to promote an image of an engaged commercial party in the fight against harmful content, creating an image that allows them to 'bluewash' their content moderation policies.[90] Secondly, engaging in co-regulation allows platforms to avoid full-fledged formal regulation, with the concomitant risk of

[85] Christina Angelopoulos and others, 'Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation' (Institute for Information Law 2016) 58–59.

[86] Jessy M Emaus, 'European Union Co- and Self-Regulation and the Protection of Fundamental Rights' in Madeleine De Cock Buning and Linda Senden (eds), *Private Regulation and Enforcement in the EU* (Hart 2020) 454–455.

[87] Giovanni De Gregorio, 'Digital Constitutionalism across the Atlantic' [2022] Global Constitutionalism 1, 18.

[88] European Court of Human Rights, *Delfi v Estonia,* judgement of 16 June 2015, 64569/09, European Court of Human Rights, *Tierbefreier eV v Germany*, judgement of 16 January 2014, 45192/ 09, see also Janneke Gerards, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023) 101.

[89] Code of Conduct on Countering Illegal Hate Speech Online https://ec.europa.eu/newsroom/just/ document.cfm?doc_id=42985; Communication on Tackling Illegal Content Online, COM(2017) 555 final, Communication Recommendation (EU) 2018/334 on Measures to Effectively Tackle Illegal Content Online, Communication on Tackling Online Disinformation, COM(2018)236 final

[90] On bluewashing, see Evelyn Douek, 'The Limits of International Law in Content Moderation' (2020) 6 U.C. Irvine Journal of International, Transnational and Comparative Law 37.

sanctions being imposed in case of non-compliance with the legislation.[91] Thirdly, co-regulatory instruments can target harmful content that platforms might want to target anyway in creating a platform that is attractive for advertisers.[92] The willingness of platforms to cooperate with co-regulatory efforts makes the EU's appeal to them very effective. Regardless of their mostly voluntary cooperation, however, the question can be asked whether the threat of full regulation can be regarded as a type of informal pressure leading to moderation practices that potentially affect the right to freedom of expression and could create tension with the rule of law.

4.3 Constitutional Concerns Accompanying Jawboning in European Speech Governance?

In Bambauer's taxonomy of jawboning, the essential rule of law tension comes into being when a government pressurizes or coerces without having the authority to do so.[93] This relates to constitutionalism–the question of whether the state has the competence to take action–and the rule of law: are the actions legal and are sufficient safeguards provided to avoid arbitrariness and fundamental rights violations? In the EU, as explored in Sect. 3, speech can be limited–even through internet intermediaries–within the confines of fundamental rights and the rule of law. This means that a restriction can be justified if it has a legal basis, serves a legitimate aim and meets requirements of suitability, necessity and proportionality. The latter requirement implies a balancing of the right to freedom of expression and the public interest that the state seeks to protect.[94]

Thus, (informally) pressurizing internet intermediaries into co-regulation can be considered a constitutional risk if the co-regulation potentially affects citizens' fundamental rights and is not sufficiently justified in light of fundamental rights requirements. When an internet intermediary complies with that pressure and starts removing content, it will seem like the internet intermediary is the accountable party for doing so. That party cannot be held accountable under freedom of expression standards in cases where content is wrongfully removed. In the meantime, the state, which is responsible for guaranteeing the freedom of expression, remains in the background, with limited transparency as to the fundamental rights risks of moderating certain types of content. The speech that is potentially wrongly removed

[91] Siva Vaidhyanathan, 'Facebook and the Folly of Self-Regulation' *Wired* <https://www.wired.com/story/facebook-and-the-folly-of-self-regulation/> accessed 14 April 2025. See also Kreimer (n 18).

[92] Gillespie (n 41).

[93] Bambauer (n 19) 88.

[94] See for example European Court of Human Rights, *MTE v Hungary*, judgement of 2 February 2016, 22947/13.

in the context of these co-regulatory documents is not protected under the safeguards of freedom of expression norms, even though the state is indirectly involved in its removal through its involvement in the co-regulatory documents. The next section illustrates this.

5. Code of Conduct on Countering Illegal Hate Speech Online

This section illustrates the abovementioned tension between jawboning and the rule of law, using the Code of Conduct on Countering Illegal Hate Speech Online ('Code of Conduct') as an example.[95] The Code of Conduct was introduced as an instrument in 2016. It was signed by the European Commission and four major internet intermediaries: Facebook, Twitter, YouTube and Microsoft. They were later joined by Instagram, Google+ and Snapchat.[96] The Code of Conduct is the result of the activities of the European Internet Forum, which is a meeting point for stakeholders, including Europol, tech companies, and EU officials.[97]

The ambition of the Code of Conduct is that companies take the lead in countering illegal speech online by *inter alia* prohibiting incitement of violence and hateful conduct in their terms of service and having a notice-and-takedown system in place that removes illegal hate speech within 24 hours. The signatories are assessed yearly, with six monitoring rounds being completed thus far. The Code of Conduct is a non-binding instrument with little oversight other than the yearly monitoring rounds, but it has resulted in 2021 in an assessment rate of content of 81%, with 62.5% of that content being removed.[98]

In 2016, the results of the first monitoring round of the Code of Conduct were evaluated critically. European Commissioner Jourova warned that 'if [tech companies] want to convince me and the ministers that the non-legislative approach can work, they will have to act quickly and make a strong effort in the coming months.'[99] This shows that state actors may use the threat of imposing formal legislation to induce strong self-regulation by social media platforms, in line with the pattern described in Sects. 4.2 and 4.3.

---

[95] Code of Conduct on Countering Illegal Hate Speech Online https://ec.europa.eu/newsroom/just/ document.cfm?doc_id=42985.

[96] Information Note Progress on Combating Hate Speech Online through the EU Code of Conduct https://ec.europa.eu/info/sites/default/files/aid_development_cooperation_fundamental_ rights/assessment_of_the_code_of_conduct_on_hate_speech_on_line_-_state_of_play 0.pdf

[97] Heldt (n 18) 60.

[98] 6th Evaluation of the Code of Conduct, 7 October 2021 https://ec.europa.eu/info/sites/default/ files/factsheet-6th-monitoring-round-of-the-code-of-conduct_october2021_en_1.pdf

[99] Clark (n 39).

The European Commission increased the pressure on the internet intermediaries to actively monitor the content on their platforms by means of the Commission's Communication on Tackling Illegal Content Online ('Communication').[100] This non-binding communication has a broader scope than the Code of Conduct, and also covers terrorist content, child pornography and copyright infringement. It focuses on the enforcement of (non-binding) EU instruments in those fields by internet intermediaries, aiming to guide them into taking more action to remove illegal content. It is non-binding, but recommends that internet intermediaries implement proactive measures such as filtering to identify and remove illegal content on their platforms, and emphasises that in cases of particularly harmful content, it may be necessary to impose timeframes to do so.[101] Such filtering obligations have been met with great skepticism when imposed by the state directly because of the risk they pose to the freedom of expression. The Communication addresses this only very briefly, however, recommending that intermediaries should introduce adequate measures to prevent over-removal.[102]

The Communication was soon followed by another EU communication: the Recommendation on Measures to Effectively Tackle Illegal Content Online.[103] This, too, is a non-binding instrument, but it is a more substantive version of the earlier Communication. The Recommendation clarifies and consolidates the efforts of the earlier documents, such as transparency and notice-and-takedown regimes. Similar to the Communication, it advises internet intermediaries to rely on automated mechanisms to detect illegal content.[104] It does not elaborate on the potential effects on freedom of expression of its recommendations, and only recommends that safeguards should be present to respect freedom of expression.[105]

The Code of Conduct and the later Communications are based on the Framework Decision on Racism and Xenophobia, which itself finds its basis in Article 29 Treaty

---

[100] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Tackling Illegal Content Online Towards an enhanced responsibility of online platforms, COM(2017) 555.

[101] Ibid., also section 3.2.

[102] Ibid., p. 6. It is worth noting that in some studies following *NetzDG,* which has a similar incentive for over-removal, the risk of over-removal seems limited, see for example William Echikson and Olivia Knodt, 'Germany's NetzDG: A Key Test for Combatting Online Hate' (Centre for European Policy Studies 2018) 2018/09; Quintel and Ullrich (n 53).

[103] Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, C/2018/1177, *OJ L 63,* 6.3.2018, p. 50-61.

[104] Ibid., rec 36, paras 18–20 also emphasising that effective and appropriate safeguards should be provided to ensure that decisions taken concerning content are accurate and well-founded.

[105] Ibid., rec 13; para 1; it could be argued that some of those possible safeguards have now been adopted in the Digital Services Act, see Articles 14 and 34.

on European Union ('TEU') on police and judicial cooperation in criminal matters.[106]

The Framework Decision approximates member states' regulation on criminalizing hate speech. Hate speech is defined as 'conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.' Member states enjoy a wide discretion in applying its provisions.[107] The wide discretion is a necessity because the determination of what constitutes hate speech is context-dependent.[108] Moreover, deciding on whether hate speech can be justifiably removed involves the balancing act explained in Sect.3. The Code of Conduct and the two Communications squarely place this balancing act with the internet intermediary, rather than the national authorities.[109] From a perspective of protection of fundamental rights and the rule of law, this creates a double problem. First, internet intermediaries may not be equipped to make the right decisions when defining hate speech and thus their act may present a risk for under- or over-removal of content, and second, it is more difficult to hold them to account for the decisions they make– regardless of the efforts made by the ECtHR and the CJEU.[110]

At first glance, it is logical that the intermediaries are not held to fully comply with the rule of law standards as they are private entities. However, it can be maintained that the European Commission itself is bound to respect the fundamental rights laid down in the EU Charter *ex* Article 51(1), also when proposing co-regulatory instruments.[111] Thus, even though private entities are not direct addressees of fundamental rights documents, the Commission has to ensure that those norms are respected in their co-regulatory efforts. This is echoed in paragraph 22 of the 2003 Interinstitutional Agreement on Better Law-Making, stating that the Commission will scrutinise self-regulatory practices under the EC Treaty.[112] The Code of Conduct is

[106] Council Framework Decision 2008/913/JHA on combating certain forms of expressions of racism and xenophobia by means of criminal law.

[107] Quintel and Ullrich (n 53).

[108] European Court of Human Rights, *Belkacem v Belgium*, judgement of 27 June 2017, 34367/14, European Court of Human Rights, *Perincek v Switzerland*, judgement of 15 October 2015, 27510/ 08.

[109] Molly K Land, 'Against Privatized Censorship: Proposals for Responsible Delegation' (2020) 60 Virginia Journal of International Law 363.

[110] David Kaye, *Speech Police: The Global Struggle to Govern the Internet* (1st edn, Columbia Global Reports 2019); Barrie Sander, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939, 960–961. European Court of Human Rights, *Perincek v Switzerland*, judgement of 15 October 2015, 27510/08, paras 204–208.

[111] Quintel and Ullrich (n 53) 210–211.

[112] It is important to note this document has been replaced by the Interinstitutional Agreement on Better-Lawmaking of 13 April 2016; that document mandates fundamental rights impact assessments for each legislative proposal, and provides that

created as a 'voluntary'[113] instrument from a discussion between stakeholders, whereas the Communication and Recommendation are unilateral communications from EU institutions. They therefore should respect the rights laid down in the Charter. This means that when facilitating the debate on this Code of Conduct, as well as taking part in the discussion of the European Internet Forum–both forms of state involvement that could be interpreted as jawboning–, the European institutions must ensure that the potential effects on fundamental rights and rule of law values are taken into account.

Second, even if such effects on fundamental rights and rule of law values are taken into account, this does not form a full solution to the problem of jawboning. It is striking that even though the Code of Conduct 'stresses the need to defend the right to freedom of expression', and the Communication and Recommendation find that fundamental rights must guide content moderation and safeguards should be foreseen to prevent collateral censorship, they nowhere expressly strike a balance between protecting those fundamental rights and achieving the policy aims of the Code, Communication and Recommendation. Such decisions are left to be made by the internet intermediaries, which, being left to their own devices to interpret fundamental rights standards, may make choices that are difficult to defend. After all, definitions of hate speech require weighing the rights of speakers against the rights of victims and the public interest. When developing and enforcing hate speech laws, member states need to account for any interference with the right to freedom of expression of the speaker or the right to the personal life of the victim: unnecessarily and unjustified impeding on either of those rights constitutes a violation of fundamental rights.[114] When attributing this task to internet intermediaries, this essentially means that private actors come to define what is hate speech, without there being an enforceable obligation to take account of the public interests or conflicting fundamental rights, and without there being adequate checks and balances or legal protection. In practice, it is often seen that internet intermediaries try to meet their obligations by means of their terms of service, but from a rule of law perspective, such terms of service might fall short–they can easily be changed and amended, even without explicit consent from internet service users,

---

explanatory memoranda of legislative instruments should explain their compatibility with human rights. It does not specify any specific requirements for co-regulation of self-regulation, however.

[113] Daphne Keller, 'Internet Platforms: Observations on Speech, Danger, and Money' (Hoover Institution 2018) 1807; Land (n 109).

[114] European Court of Human Rights, *Altug Taner Akam v Turkey*, judgement of 25 January 2012, 27520/07, European Court of Human Rights, *Selahattin Demirtas v Turkey*, judgement of 22 December 2020, 14305/17.

none of the platforms have copied the definition laid down in the Framework Decision, and there is hardly any control of their contents.[115] Indeed, this is a matter that is not sufficiently addressed by the case law of the CJEU and the ECtHR so far. Although these courts have made clear that, to a certain extent, internet intermediaries can be held responsible in the courts for removing or not removing certain content, this case law does not address rule of law related matters such as regulation through the terms of service, nor does it address the role state pressures play in this.

Hence, the Code of Conduct and the Communications clearly show how informal pressure to cooperate with co-regulation creates fundamental rights and rule of law risks. EU institutions leave the definition of legal speech to internet intermediaries, creating risks for over- and under-removal of content, as well as legal uncertainty and a lack of legal protection. It therefore can be maintained that by informally pressuring internet intermediaries to adopt certain moderation policies, the EU does not sufficiently respect fundamental rights itself.

Conclusion

This chapter set out to answer the following question: from a rule of law perspective, what are the constitutional fragilities of jawboning in European co-regulatory instruments concerning speech governance by internet intermediaries? It has explained that jawboning can be at odds with constitutional and rule of law values when state actors exercise informal pressure on private entities in the pursuit of a policy aim that may (result in) infringing the right to freedom of expression. Firstly, there is a risk of sidestepping constitutional safeguards around speech limitation. By relying on private entities to moderate harmful content, governments can indirectly restrict speech acts without there being a clear balancing of fundamental rights for which the state itself can be held responsible. Rather, the internet intermediary is responsible, but not held to the same constitutional framework for protecting and banning certain types of content.

Secondly, co-regulation makes internet intermediaries *de facto* regulators, who have to provide definitions and identify instances of hate speech or other types of unlawful content on their own. Laying down such definitions in terms of service puts significant strain on the principles of legal certainty and checks and balances, as well

[115] Natalie Alkiviadou, 'Hate Speech on Social Media Networks: Towards a Regulatory Framework?' (2019) 28 Information & Communications Technology Law 19, 19; Daniel Konikoff, 'Gatekeepers of Toxicity: Reconceptualizing Twitter's Abuse and Hate Speech Policies' [2021] Policy & Internet 502.

as creates a certain potential for under- and over-removal of content.

Thirdly, some forms of enforcement, such as excessive filtering, have been shown to pose a risk to freedom of expression and information. General monitoring is prohibited under EU law for that reason. However, the CJEU and other EU institutions have held that internet intermediaries sometimes can be obliged to deploy automated filtering mechanisms to combat unwanted types of speech, which can be seen as an indirect pressure to employ general monitoring practices. This is problematic from a rule of law perspective and poses a threat to freedom of expression due to the risk of collateral censorship.

Pursuing policy aims through informal pressure is attractive, as it allows state actors to act fast and effectively by targeting internet intermediaries, who have the resources to combat harmful content. Discussing policy aims with stakeholders like internet intermediaries also includes their potential cooperation, and most internet intermediaries have an incentive to cooperate and take an active stance against harmful content–even if only to avoid proper regulation. It, therefore, is important to look for ways to ensure that these advantages can be realised while minimising the risks for the rule of law and fundamental rights violations as identified in this chapter. One way to do so would be to ensure that each of the instruments resulting from, or with the intent of, jawboning contains a sound explanation of legal basis, a sufficiently clear definition of what types of speech are concerned, as well as laying down standards to inform a balancing of interests when freedom of expression is concerned for the internet intermediary. In particular, it should be explained how the internet intermediaries' content moderation can be considered as a suitable, necessary and proportionate means of pursuing a legitimate aim in the public interest. This would better reflect the fundamental rights at stake in content moderation than the simple 'sufficient safeguards should be in place' that current co-regulatory instruments state. In terms of redress, it would make sense that co-regulatory instruments provide a clarification of the responsibility each actor takes in case a co-regulatory instrument leads to an infringement of fundamental rights. The civil redress opportunities that internet users have vis-à-vis the internet intermediary do not fully reflect the role of the state in the content moderation actions of the internet intermediary. Finally, to ensure better compliance with the rule of law, recommending general monitoring informally should be avoided: Article 8 DSA prohibits general monitoring obligations, and informal recommendations with varying levels of coercion can be construed as obligations.

Afterthoughts

The main argument of this contribution has been that co-regulation can exert pressure on internet intermediaries to moderate content. If this leads to content moderation that affects legally acceptable speech because it is aimed at unclearly defined speech acts, it risks interfering with the freedom of expression. This is problematic, since influences in co-regulatory processes are often opaque, meaning that any interferences with the freedom of expression cannot be scrutinised along the requirements for limitations set out in article 10(2) ECHR and 52(1) CFR. The contribution has used the Code of Conduct on Countering Illegal Hate Speech as an example. However, since the finalisation and publication of this contribution, the DSA's framework for codes of conduct has become operational, warranting an addendum on recent events.

The Digital Services Act introduces three specific grounds for regulators to encourage and facilitate the drawing of voluntary codes of conduct: for tackling illegal content or systemic risks,[116] for transparency in advertising chains,[117] and for promoting accessibility.[118] Currently, the Code of Conduct on Countering Illegal Hate Speech Online+ (signed by 12 platforms) has been adopted into the framework of the DSA under article 45, as has the Code of Conduct on Disinformation (40 signatories). The text of article 45 emphasises that codes of conduct are seen as 'voluntary', similar to the Code of Conduct examined in Section V. The question remains how 'voluntary' the encouragement and facilitation of the drafting of such codes of conduct under the DSA is. The DSA stresses, on the one hand, that signing a code of conduct does not lead to the presumption of compliance, but on the other hand it emphasises that refusal without proper explanation to participate in a code of conduct can be taken into account when establishing non-compliance.[119] It is evident from this phrasing that there is significant pressure emanating from the Commission and regulators for online platforms to participate in codes of conduct. This may be problematic when participation in a code of conduct leads to the removal of legal speech. This is in line with Griffin's critique that labelling such codes of conduct as 'voluntary' is used to mask "highly politicised questions of online speech governance."[120] This can be a particularly salient point of criticism in

[116] Regulation 2022/2065, art. 45.
[117] Regulation 2022/2065, art. 46.
[118] Regulation 2022/2065, art. 47.
[119] Regulation 2022/2065, rec. 104.
[120] Rachel Griffin, 'Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns' (2024) 2024 Technology and Regulation 167, 185.

the case of the Code of Conduct against Disinformation, as there can be a risk of internet intermediaries being incentivised to remove legal speech.[121]

As such, the DSA forms a two-sided sword with respect to the argument of this contribution. On the one hand, it provides clear legal bases for regulators to engage in encouraging and facilitating the drafting of codes of conduct, and therefore creates a degree of transparency and accountability around the practice of co-regulation. On the other hand, it provides a legal incentive for internet intermediaries to participate in codes of conduct to avoid being seen as non-compliant with the DSA, thus providing an almost formal coercion to participate. The codes of conducts so far have tackled widely defined types of illegal speech. They may therefore risk inadvertently promoting the removal of legal speech by participating internet intermediaries. The DSA therefore provides a 'mixed bag' solution, and does not present a full answer to the criticism made in this contribution.

[121] Aleksandra Kuczerawy, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?' in Georgios Terzis and others (eds), *Disinformation and Digital Media as a Challenge for Democracy* (1st edn, Intersentia 2019) <https://www.cambridge.org/core/product/identifier/9781839700422%23c15/type/book_part> accessed 2 October 2025.

Chapter VIII: Conclusion

The Digital Services Act aims to order a chaotic landscape of platform governance. The landscape is chaotic from a regulatory perspective as various member states have introduced requirements beyond the framework of the ECD, and rules of jurisdiction on the internet are generally unclear. It is also chaotic from a practical perspective, as internet intermediaries moderate in different ways, and interact with various, differing stakeholders such as users, state actors, and civil society. Within this chaotic landscape, the EU regulator strives to harmonise those aspects of information society services that it deems most important for safety and economic prosperity in the digital realm. The DSA requires that internet intermediaries provide notice-and-action mechanisms, options for recommender systems, transparency on the provision of their services, and, in cases of VLOPSEs, clarification of how they bear responsibility for the potential societal impacts of their services. The DSA also regulates who interacts with the content moderation process, and how. These aspects of the DSA determine how content is moderated, and therefore, by extension, how the freedom of expression on the internet is shaped. While these procedures create a sense of order in the realm of social media, they also raise questions. How are interests of various actors balanced in this procedural framework? What effects does that have on the right to freedom of expression of internet users? And, who conducts that balancing act?

While many of these questions have been addressed outside of the framework of the DSA, they have not (yet) been explored exhaustively within the DSA. This dissertation forms a starting point for a discussion on these questions, by providing an analysis about the implications for freedom of expression under the DSA's procedural framework for content moderation and the actors it involves. The introductory chapter illustrated that the governance of internet intermediaries is not an affair that involves solely state actors, or solely the internet intermediary. Instead, the governance is in the hands of a web of actors that interact with one another, represent different interests, occupy a different information position, and have different levels of power in the content moderation process. The introductory chapter has used Caplan's networked platform governance theory to describe such interactions as a governance network. The DSA creates a procedural framework that resembles a governance network of actors involved in content moderation, all interacting, but with different interests and different power positions. This network ultimately shapes the freedom of expression on the internet.

The DSA resolves questions posed to the freedom of expression by referring to the Charter, and, indirectly, to the European Convention on Human Rights. In spite of these general references, it is not always clear how the freedom of expression of users is affected by the DSA's provisions. To create clarity on that framework, this dissertation has assessed the question how the DSA and the formal roles for various actors in content moderation affect the balancing of freedom of expression of internet users. It has endeavoured to answer this question for three classes of actors in the framework of networked platform governance in the DSA: internet intermediaries, civil society, and (supra-)state actors. This concluding chapter provides a brief overview of the findings presented in the previous chapters, and subsequently answers the overarching research question. This is followed by a brief synthesis of the suggestions for improvement made in the previous chapters, resulting in a call for more accountability for balancing the freedom of expression in content moderation in the DSA's networked platform governance.

1. The Internet Intermediary in Networked Platform Governance

The first subquestion addressed in this dissertation is how the role of the internet intermediary and its interactions with other stakeholders in content moderation in the DSA affects the balancing of freedom of expression of internet users. This dissertation has examined this question by assessing how internet intermediaries provide transparency, and assessing how internet intermediaries can be held liable for hosting illegal content in the DSA. Chapter II functioned as an empirical introduction into content moderation, describing who moderates content, and on the basis of what. This empirical basis forms a foundation for observations about content moderation practices in subsequent chapters. Chapter II explored the DSA transparency database as a novel instrument with a twofold purpose: to form an empirical description of content moderation as reported by online platforms, and to scrutinise how the transparency database contributes to transparency about content moderation processes. Following article 24(5) DSA, online platforms are required to upload statements of reasons for content moderation decisions in a publicly accessible database. This database informs other stakeholders in the networked platform governance framework. The theoretical framework of Chapter I thereby highlighted the different information positions that actors in a governance network can have, due to the ability of data providers to determine the scope of- and discourse on the information that they provide. The information provided by internet intermediaries determines the degree to which they can be held accountable for their practices. Therefore these databases form what Leerssen calls 'accountability interfaces', since they create automated ways to disclose non-

sensitive data.[1] The chapter concluded that non-sensitive data on content moderation can not only be used to hold platforms accountable for their content moderation practices, but can also foster content moderation practices because of the potential for public backlash in case such transparency indicates poor content moderation practices.

The DSA transparency database illustrates that content moderation decisions are made at unprecedented scale and speed, with millions of decisions being uploaded every day. The majority of content moderation takes place outside of the realm of social media platforms: most decisions are uploaded by e-commerce platforms. The statements of reasons contained in the database show that power in content moderation is centred with the internet intermediary: most of the moderation decisions are taken at the initiative of the internet intermediary, detected and decided automatically, and decided based on terms and conditions drafted by the internet intermediary. This confirms the notion that internet intermediaries act as sovereign actors in the digital realm. Only a small minority of decisions is taken based on notices submitted by users and trusted flaggers.

Chapter II indicated that, by and large, the DSA transparency database architecture is compliant with the legal requirements of article 17 DSA, and could in principle help provide transparency on content moderation practices. However, the chapter also showed that internet intermediaries currently do not use the transparency database to provide meaningful insights to other stakeholders into content moderation beyond aggregate data. Internet intermediaries predominantly report removing content based on vaguely formulated categories, which do not provide specific insight into the problematic nature of the content they removed. This lack of transparency does not enable scrutiny by regulators or civil society of platform practices, and therefore impedes the accountability promised by this 'accountability interface'. Transparency amongst stakeholders in networked governance models is necessary for effective cooperation but is currently not supported by the submissions online platforms make in the transparency database.

Such scrutiny, and therefore accountability, is facilitated at the individual level by the conditional intermediary liability exemption framework for hosting providers that has formed the cornerstone of platform regulation in the European Union for the past 25 years. Chapter III argued for expanding possibilities for scrutiny by

[1] Paddy Leerssen, 'Seeing What Others Are Seeing' (University of Amsterdam 2023) 183–184.

narrowing the scope of intermediary liability exemption to exclude bad faith actors. Liability exemption was a core part of the e-Commerce Directive (specifically articles 12-14). To stimulate economic activity and protect freedom of expression, hosting providers are exempted from liability for hosting illegal content if they are not aware of that content or act expeditiously once learning about it. Whether a hosting provider is deemed aware of illegal content is determined by the notion of the 'diligent economic operator', a standard that has been interpreted narrowly in case law of the Court of Justice of the European Union (*Google France, L'Oreal, Sotiris Papasavvas, Youtube/Cyando)*. Essentially, hosting providers are not aware of illegal content unless they are alerted by a sufficiently precise and adequately substantiated notice submitted by concerned parties. This doctrine has been included almost *verbatim* in the DSA (specifically articles 4-6). However, the internet is a very different place now than it was 25 years ago when the notion was originally adopted. Applying the same doctrine as 25 years ago again may not be desirable. The chapter therefore argued for using provisions in the DSA to guide a narrow interpretation of the liability exemption. This narrow interpretation can achieve a fairer burden of moderation, and better reflects developments around content moderation capabilities, the need for increased platform responsibilities, and the higher standard of diligence required by the DSA and sectoral regulation.

Although the freedom of expression is often raised as a counterargument to a narrow interpretation of the liability exemption, the case law of the European Court of Human Rights does not necessarily preclude such an interpretation. The purpose of a narrower interpretation of liability exemption would be to exclude 'bad faith' hosting providers from enjoying the liability exemption, and thus could incentivise them to take more responsibility for their content moderation practices. Legal bases for such exclusion can be found in article 16 and 23 DSA. Article 16 requires hosting providers to facilitate a functioning notice-and-action mechanism. The notice-and-action mechanism forms the very core of the intermediary liability exemption, and failing to facilitate these mechanisms is a bad faith strategy to avoid potential liability. Exclusion from the scope of the intermediary liability exemption is therefore warranted. Similarly, article 23(1) DSA requires hosting providers to suspend users who frequently upload manifestly illegal content. The expectation to remove content is borne from the fact that the frequency of uploads and manifestness of their illegality is such that it requires any good faith diligent economic operator to suspend services to an account disseminating such content. When failing to do so, this can be seen as acting in bad faith and therefore warrants exclusion from the intermediary liability exemption.

Furthermore, it has been noted that the intermediary liability exemption was originally devised to protect the freedom of expression by not risking over-removal of content as a result of liability. However, as remarked above, the balance may have shifted in the 25 years since it was developed. As the existing liability exemption framework can also apply to bad faith actors, illegal content may be under-moderated. Sectoral regulation, such as the Copyright in the Digital Single Market Directive (Directive 2019/790)[2], acknowledges this development, but it has not found its way into horizontal regulation. This means that the interests of those aggrieved by infringing content may not be effectively weighed against bad faith hosting providers' or users that post manifestly illegal content. As a result, it may hamper an adequate balancing of the freedom of expression against other interests and rights.

Excluding bad faith actors from the liability exemption will require hosting providers to take more responsibility in platform governance. It protects the right to freedom of expression by incentivising more responsibility for a good faith stance toward content moderation, ultimately helping the public debate. This argument does not depart from the original premise of conditional liability exemption: good faith hosting providers will remain protected under article 6 DSA, and will not be forced to monitor content in the sense of article 8 DSA. Therefore, the narrowing of the liability exemption should not have a chilling effect on freedom of expression. It merely strengthens responsibility for platform governance by holding bad faith hosting providers liable in clearly delineated circumstances under the DSA.

2. Users and Civil Society in Networked Platform Governance

Notice-and-action mechanism form the predominant avenue through which users can signal illegal content to hosting providers. These mechanisms form a core component of the intermediary liability framework, not in the least because they are an important means of interaction between user and hosting provider in networked platform governance. Users use notice-and-action mechanisms to report illegal content and contribute to platform governance, and as such play a role in content moderation. Notice-and-action mechanisms empower users to take ownership over the public debate, and protect their right to information. However, the degree to which users are empowered by notice-and-action mechanisms is dependent fully on how those mechanisms are facilitated by internet intermediaries. Article 16 DSA creates a procedural framework around notice-and-action mechanisms. How

[2] Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, *OJ L 130,* 17 May 2019, p. 92-125.

internet intermediaries comply with article 16 DSA therefore affects the balancing of the freedom of expression against other rights and interests in content moderation.

Chapter IV has investigated this framework by conducting a walkthrough of notice-and-action mechanisms of four platforms. Article 16 DSA requires hosting providers to facilitate easily accessible and user-friendly mechanisms to report illegal content. These mechanisms trigger the conditional liability exemption framework explored in Chapter III: if a notice is submitted through these mechanisms, a hosting provider is expected to remove content expeditiously if it is illegal. The walkthrough shows that notice-and-action mechanisms take different forms, depending on the type of service offered. Their designs are also shown to affect how users report content: some designs are likely to complicate reporting illegal content, or even discourage reporting illegal content by hiding reporting functionalities. This is not only incompliant with article 16, but also creates a tension with the intermediary liability exemption. After all, if there is no easily accessible or functional mechanism to create awareness with the hosting provider about illegal content, there are limited means of triggering the conditional liability mechanism.

Chapter IV has also shown that failing to provide such mechanisms diminishes the role of users in networked platform governance. Indeed, users are only as empowered as the platform allows them to be. This does not directly affect how users are able to express themselves on the internet; however, it does affect the way in which they can interact with the process of content moderation, the degree to which they affect how others express themselves on the internet, which does ultimately shape the freedom of expression. A functioning notice-and-action mechanism allows users to signal to platforms how they want to shape the freedom of expression on the internet. It also provides a remedy for aggrieved parties to balance their interests against the right to freedom of expression of another party; this allows them to take an active stance in shaping the freedom of expression on the internet.

The representation of civil society in content moderation is organised in part through the participation of trusted flaggers. Trusted flaggers have been a common presence in platform governance for some time. Most large internet intermediaries offer privileged reporting mechanisms to civil society organisations, governments, and sometimes individuals. Although the involvement of trusted flaggers takes different forms, it usually involves interacting with notice-and-action mechanisms to report illegal- or non-compliant content. Article 22 DSA sets out a structure for this: independent civil society organisations can seek to be certified by their national

digital services coordinator as a trusted flagger and thus formally become a part of the platform governance network. Once certified, trusted flaggers report content through the mechanisms of article 16. Those reports must be treated with priority. Chapter V has shown that the formal role for trusted flaggers in the DSA affects the balancing of freedom of expression in content moderation, and must be surrounded by adequate safeguards.

Chapter V explored the dynamic around accreditation of trusted flaggers and their subsequent reporting practices, as the lack of voluntariness in this framework has implications for the freedom of expression of internet users. Trusted flaggers represent specific interests, which do not need to be balanced against the freedom of expression of other internet users. At the same time, their flags are taken to carry more weight than ordinary reports based on the perceived expertise of the trusted flagger. Consequently, hosting providers may be more inclined to remove flagged content without careful review, which poses a risk to freedom of expression. This risk is manifest especially in light of the specificity of the interests that trusted flaggers represent. National interests and preferences can have an overbearing effect through this dynamic: trusted flaggers are certified by national digital services coordinators, but their activities have an effect across the European Union. In cases of national supervising authorities being appointed as trusted flagger, the risk to freedom of expression is even more pertinent.[3]

However, trusted flaggers can also help online discourse through representing interests that would otherwise be overlooked in the content moderation process, or because civil society actors and oversight authorities have a fast-track way to signal their concerns with content. Their participation in networked platform governance is therefore desirable, and needs to be balanced against the freedom of expression risks that privileged flagging presents. In order to effectively fulfil their purpose in content moderation, trusted flaggers require resources. Funding of such civil society organisations is simultaneously crucial and legally contentious: they have a requirement of independence, meaning funding by social media platforms is not allowed, even though they would be the logical funder of such organisations considering their economic power and them being the ultimate beneficiary of trusted flagger services. Further, the Commission has explicit guiding power under article 22(8) to instruct DSCs on the selection of such actors, and should use this

[3] More specifically on this: Jacob van de Kerkhof, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' in José van Dijck and others (eds), *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values* (Amsterdam University Press 2025).

competence to clarify requirements of objectivity and independence. An important safeguard for the freedom of expression of users would be the addition of a correction mechanism for instances when a national digital services coordinator appoints abusive or malevolent trusted flaggers. Platforms have the right to suspend their flagging privileges under article 23(2), but a remedy should also be available within the structure of the Board of Digital Services Coordinators to appeal the appointment of trusted flaggers in other jurisdictions and urge the competent DSC to withdraw certification of a malevolent trusted flagger. This ensures accountability within this aspect of networked platform governance.

3. The State in Networked Platform Governance

In networked platform governance, users, civil society, and internet intermediaries, are dependent on state- and supra-state regulators enforcing the DSA. Indeed, although the attribution of competencies in content moderation and platform governance to private entities indicates a reduced role of the state, this need not be the case. Freeman has argued that the role of the state is shifting, but not weakening; instead of acting strictly as regulator, states facilitate negotiations between stakeholders, and can even act as a stakeholder themselves.[4] This is visible – to some degree – in the DSA: newly attributed state entities act as a facilitator in regulating and sanctioning platform content practices under the DSA. The way in which they supervise and enforce the DSA shapes the balance struck between the freedom of expression and other rights and interests in content moderation, as has been explored in Chapter VI. The role of state actors as stakeholders has been explored in Chapter VII, which concentrated on the implications for freedom of expression when states enter negotiations between platforms and other stakeholders.

Chapter VI has described the role national regulatory entities take in relation to the freedom of expression of internet users. Digital Services Coordinators (DSCs) hold powers that interact with content moderation, which can affect the scope of freedom of expression of users. They can directly interact with users' content, certify stakeholders in networked platform governance, or sanction content moderation practices. Similarly, the degree to which they effectively enforce the DSA determines how users' rights are protected under the DSA's regime, which is partly aimed safeguarding users' fundamental rights. In short, they act as the conductor of the networked platform governance orchestra – and how they conduct the orchestra determines the symphony of the freedom of expression.

[4] Jody Freeman, 'The Private Role in the Public Governance' (2000) 75 New York University Law Review 543, 671.

DSCs are attributed competence in a novel, hybrid structure. This structure is based on the GDPR's one-stop-shop model, with the important connotations that a) DSCs are not necessarily independent administrative entities, b) they are part of a non-binding cooperative framework, and c) power in the enforcement network is partly centralised with the European Commission for the enforcement of VLOPSE-specific provisions of the DSA. The strict national competencies of DSCs in the one-stop-shop model run the risk of national enforcement authorities applying national law extraterritorially, by enforcing the DSA along national constitutional and legal traditions and values. The exclusive competence DSCs have in their own territory precludes correction by other DSCs or the Board of Digital Services Coordinators. The complex enforcement structure causes concern regarding the protection of user's rights in this framework: networked enforcement of previous regulations in a one-stop-shop model has shown that this can administratively paralyse the enforcing authorities that have comparatively the largest case load, while simultaneously giving those authorities significant power in shaping the enforcement of European regulation. Not only is the risk for paralysis present, the different nature of authorities acting as DSCs can also create disparity in DSA enforcement: media authorities might look differently upon enforcement powers than competition authorities, which again might affect how user's rights are protected in the DSA's framework. Chapter VI therefore proposed that the Commission should provide further guidance to anchor the protection of freedom of expression in DSA enforcement, and take the lead in cases where national enforcement creates risks for the freedom of expression of users.

Chapter VI focused on the role of national enforcement authorities as guardians of the freedom of expression of internet users against the threats that internet intermediaries pose. In contrast, Chapter VII theorised that in informal interactions with internet intermediaries regulators pose a risk to freedom of expression. To interrogate this dynamic, it adopted the lens of 'jawboning', a term coined in predominantly U.S. legal scholarship, to describe the dynamic of informal pressure being exerted on internet intermediaries to abide by the regulator's wishes without there being an explicit legal basis to do so. It is stipulated that, in fear of regulation and stringent enforcement that jeopardises commercial interests, online platforms likely yield to informal pressure and remove content.[5] This creates a 'censorship-by-

[5] Recent trends of overt disobedience with EU regulators somewhat undermine the theory's premise. See e.g. Jacob van de Kerkhof, 'Musk, Techbrocracy, and Free Speech' (*Verfassungsblog*, 17 January 2025) <https://verfassungsblog.de/musk-techbrocracy-and-free-speech/> accessed 24 March 2025.

proxy' situation in which the platform limits speech at the behest of the state. The theoretical framework around jawboning has been developed in the United States, where interactions of regulators with internet intermediaries aimed at content removal are often found to be unconstitutional. By contrast, it is submitted that in the European Union there is a higher tolerance for such interactions.

However, informally pressuring internet intermediaries to remove content can also be problematic under the framework of article 10 of the European Convention on Human Rights. When regulators partake in the formulation of codes of conduct together with industry actors, as is recently the case for the Code of Conduct against Illegal Hate Speech Online+[6] and the Code of Conduct on Disinformation[7] in the DSA scheme of article 45,[8] they exert influence and pressure on the drafting of such codes. This is part of the dynamic described as networked platform governance in this dissertation. In acting in line with these co-regulatory instruments, internet intermediaries may well remove user-generated content. This would create a censorship-by-proxy scenario: a regulator wants a specific type of content removed, and the internet intermediary feels compelled to remove that content, resulting in an interference with the user's right to freedom of expression. Still, such would not be considered problematic if the requirements for interferences with the right to freedom of expression of article 10(2) ECHR were met, being that interferences must be provided for by law, serve a legitimate aim, and be necessary in a democratic society. However, in cases where participation with co-regulatory processes is unclear or not openly acknowledged, it is unlikely that regulators can be held accountable for non-compliance with the requirements for article 10(2) ECHR, however noble the purpose of their involvement in the drafting of these codes of conduct. Opacity around co-regulatory processes therefore may breach the freedom of expression of internet users. Regulators therefore should be encouraged to be transparent about their involvement, which would allow scrutiny over the role they take in this aspect of networked platform governance.

[6] The Code of Conduct on Disinformation, accessible via: https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation

[7] The Code of Conduct on Countering Illegal Hate Speech Online +, accessible via: https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

[8] Regulation 2022/2065, art. 45. The impact of these Codes of Conduct is underlined by the fact that they serve as a compliance instrument in systemic risk assessments, meaning that non-compliance with these instruments can be seen as a contra-indication for DSA compliance, see rec. 104.

Synthesis: How does networked platform governance in the DSA affect the balancing of freedom of expression in content moderation?

This dissertation has demonstrated how networked platform governance in DSA involves a range of stakeholders that influence how content is moderated, and how freedom of expression is balanced against other rights and interests. These stakeholders collectively contribute to the governance of a commercial product with a public function that shapes the freedom of expression.[9] The individual contributions which form part of this dissertation unpacked the procedures by which various actors interact with content moderation and platform governance under the DSA. The current section synthesises those findings and provides an introduction into the next section that suggests improvements in accountability in the DSA's networked platform governance.

The interactions between actors in networked platform governance shape the freedom of expression of internet users. In this process, speech can be restricted. As demonstrated, users and civil society organisations may urge internet intermediaries to remove content, or intermediaries may choose to do so in pursuit of commercial or ideological interests. These interactions may limit the freedom of expression of internet users. Ordinarily, such restrictions must comply with the requirements for legitimately interfering with the freedom of expression contained in article 10(2) ECHR and article 52(1) CFR. This means that, when enforcing the DSA, state actors are bound by the Charter and the Convention. Chapter VI has explained that in practice, however, enforcement of the DSA ensuring that restrictions of the freedom of expression meet the requirements of the Charter and the Convention may be constrained by the specific enforcement network of the DSCs. While the Convention and the Charter provide a common core for a European understanding of the scope of the freedom of expression, what is considered illegal speech in member state may differ considerably. However, in the DSA's procedural framework, it is possible that national laws and traditions are enforced EU wide. This is especially true due to the centralisation of enforcement power around a few larger Digital Services Coordinators – dubbed in Chapter VI as the 'Dublin effect'. In addition, the centralisation of enforcement presents a risk to the DSA's function in protecting freedom of expression vis-à-vis platforms. Similar enforcement structures indicate

[9] It is beyond the scope of this dissertation to treat the analogies for online platforms as public squares, public utilities, or public spaces. See on that for example: David E Pozen, 'Straining (Analogies) to Make Sense of the First Amendment in Cyberspace' in David E Pozen (ed), *The Perilous Public Square* (Columbia University Press 2019); Claudia E Haupt, 'Platforms as Trustees: Information Fiduciaries and the Value of Analogy' (2020) 134 Harvard Law Review Forum 34.

that the risk for paralysis is not insignificant, which means that complaints may go unhandled and users may be left with limited protection of their rights under the DSA.

Chapter VII illustrated another risk from the perspective of article 10(2) ECHR and 52(1) CFR: the use of informal regulation to influence content moderation practices – a phenomenon often referred to as 'jawboning'. State actors need to ensure that their conduct in networked platform governance sufficiently balances their interests in regulating the internet against the freedom of expression of internet users. Even though the motivation behind regulatory interference may be legitimate, the informality and opacity of such dealings means that there is no way of scrutinising state influence in the content moderation process, as Chapter VII has argued.

When an internet intermediary, as a private actor, decides to remove content, this creates a different tension with the Convention and the Charter. *Sensu stricto* the norms of the Charter and Convention do not directly apply to private actors, such as internet intermediaries and internet users. As explored in Chapter I, fundamental rights may still apply because regulation compels private actors to respect them and because fundamental rights are used to interpret open-ended norms in national law. Even so, decisions shaping the freedom of expression are then made by internet intermediaries. This creates friction, since internet intermediaries do not always have a clear incentive to align their content moderation policies with the law, which may result in illegal content remaining unmoderated, and legal content being moderated. This seems to call for additional procedures and mechanisms to safeguard individual rights in the digital realm.

The fact that content moderation practices do not always align with the law may cause grievances to parties affected by unmoderated illegal content. Parties aggrieved by illegal content generally suffer from the disadvantage that internet intermediaries enjoy a wide exemption from liability. Although there are remedies available to request intermediaries to remove content, most notably through notice-and-action mechanisms discussed in Chapter IV, the strict interpretation of the liability exemption may make requesting damages difficult, even in cases of bad faith actors. Chapter III has argued that a narrower interpretation of the liability exemption aligns better with the diligence expected from internet intermediaries and may therefore provide additional remedies to aggrieved parties. This would alter the balance of interests between users and intermediaries, since the interests of users affected by illegal content need to be given more weight by the platform as the risk of liability may increase. Of course, it is still necessary to consider unwanted

limitations to the freedom of expression that may result from the over-moderation that liability may cause. The narrow interpretation is therefore intended solely for internet intermediaries acting in bad faith in relation to clearly articulated diligence norms in the DSA and sectoral regulation.

By contrast, some other provisions in the DSA are aimed at shaping a procedure that allows the users to affect the balancing of freedom of expression by the internet intermediary, in a similar fashion as it would protect them against a state actor. For example, the DSA stresses the protection of fundamental rights, particularly the freedom of expression, in the application of terms and conditions by internet intermediaries, for example in situations in which content is removed.[10] Users can signal what content they want removed from a platform by interacting with a notice-and-action mechanism; this gives them ownership over their freedom of information and other users' freedom of expression. It allows them to balance their own, or societies', interests against the freedom of expression of another user. How these mechanisms are designed shapes how such balancing interactions take place, and therefore the platform determines to what extent users have ownership over the balancing of the freedom of expression. Empirical findings, set out in Chapter IV, show that some platforms design their mechanisms in such a way that they dissuade users from interacting with them. This negatively affects the ability to influence the balancing of freedom of expression in networked content moderation.

Notice-and-action mechanisms are also crucial for trusted flaggers, explored in Chapter V. Trusted flaggers enjoy a privileged position with the internet intermediary, meaning that their notices are treated with priority and they can be considered a more influential actor in content moderation than ordinary users. However, they are subject to an extra layer of requirements, as a state actor determines in what capacity organisations seeking to be selected as trusted flaggers can interact with internet intermediaries. Moreover, the interactions with trusted flaggers are a balancing act between the interests of the internet intermediary, the trusted flagger, the state actor selecting the trusted flagger, and the user whose content is subject to complaints by the trusted flagger. How this balancing act unfolds ultimately determines how the freedom of expression of users is weighed against the interests of the trusted flagger and its certifying authority. State actors may pursue their own interests in certifying trusted flaggers, and trusted flaggers may also prioritise their policy aims over fundamental rights. This means that, if

[10] Regulation 2022/2065, art. 14(4), rec. 22.

internet intermediaries do not provide a sufficient safeguard in checking notices submitted by trusted flaggers to ensure that they flag illegal content in a manner that is coherent with the Charter, freedom of expression may be under pressure.

In sum, this dissertation has addressed the question of how the DSA and roles therein for internet intermediaries, users, civil society, and state actors affect the balancing of the right to freedom of expression to other rights and interests in content moderation. It has provided an account of how the DSA creates a web of actors that interact with content moderation, and all, in their own way, affect how content is moderated. This has been shown to result in a delicate constellation of different, but interdependent interests, in which the freedom of expression of internet users may not always be appropriately weighed, or even overlooked. Even though the DSA can in some ways be seen as a major step forward in regulating digital services, it is necessary to be mindful of the freedom of expression implications of the governance network that it constructs. To support that conclusion with some policy suggestions, the next section summarises findings in the individual chapters on how accountability for the right to freedom of expression can be created within the framework of the DSA.

Toward accountability in the DSA's networked platform governance

Although the primary aim of the dissertation has been to describe how the DSA and its actors affect the balancing of freedom of expression in content moderation, at several points, suggestions and recommendations have been made to ensure that freedom of expression is adequately balanced in networked platform governance. This subsection summarises and synthesises these.

The recommendations made across the chapters can be seen as directed toward creating accountability for how freedom of expression is balanced in content moderation in the DSA's interventions in that process. To frame the recommendations around the concept of accountability, Bovens' definition of accountability can be recalled as it was explained in Chapter I: "a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences."[11] This maps well onto the notion of the DSA as a regulation that creates relationships and procedures between actors that can be described as networked platform governance, which may affect the freedom of

[11] Mark Bovens, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 European Law Journal 447, 450.

expression of internet users. The forum in this case is the network shaping platform governance. The result of platform governance is the answering of questions on what– and how content is moderated. Actors are under an obligation to explain and justify their conduct with regard to the freedom of expression of other participants in that forum. In networked platform governance, internet intermediaries can be questioned for that conduct by participants through notice-and-action mechanisms, or through public accountability on the provided transparency. This can be followed by consequences, such as liability or sanctions by oversight authorities. In networked platform governance, it is not always clear what the relative influence is of actors in the governance framework. Subsequent chapters have presented several solutions to the accountability issues this bring along.

Chapter II has described how accountability in networked platform governance can be created by internet intermediaries, which could be achieved by obliging them to inform regulators, users, and civil society of their moderating practices. In addition, the chapter argued that platforms must be more accountable for how they provide that information, which is currently incompliant with the objectives of the DSA transparency database. Chapter III argued that a narrower liability exemption could strengthen the accountability of internet intermediaries to other actors in networked platform governance. In line with Bovens' definition for accountability[12]: liability can be seen as legally institutionalised consequences for behaviour in an accountability network.[13] In the context of networked platform governance, this means the liability of the internet intermediary. Despite the increased importance of internet intermediaries in day-to-day affairs, they have not become more accountable for their content moderation practices in the form of liability. This accountability gap could be bridged through providing aggrieved parties with increased access to the legal remedy of holding internet intermediaries liable for hosting illegal content. The narrowing of the scope of the liability exemption that this requires may be justified in instances where the internet intermediary is acting in bad faith. Narrowing the liability exemption can therefore strengthen the accountability of internet intermediaries.

Chapter IV described how users can question the content moderation practices of internet intermediaries: through notice-and-action mechanisms. Such mechanisms

[12] Ibid.

[13] Jonathan GS Koppell, 'Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"' (2005) 65 Public Administration Review 94, 96.

contribute to the accountability of internet intermediaries in two ways. First, internet intermediaries can be held accountable for respecting notices submitted by users, or they risk liability for hosting illegal content. Second, they are accountable for providing the avenue through which users can question their content moderation practices. The chapter argued that not all internet intermediaries enable this form of accountability, which may hamper the degree to which they can be held accountable. It might even affect liability as a legally institutionalised form of accountability. This in turn means that users have less ownership over the content moderation process and the shaping of freedom of expression online. To frame this in terms of a recommendation: in addition to strengthened accountability through a stricter liability regime as proposed in Chapter III, the accountability of internet intermediaries for facilitating interactions with users in the content moderation process can be strengthened through stricter enforcement of the norms laid down in article 16 DSA.

Trusted flaggers are state-certified actors in content moderation with a complex position in networked platform governance. Chapter V has explored how the position of trusted flaggers in the DSA can affect the freedom of expression of internet users. For their conduct in networked platform governance, they are accountable to the certifying state actors, internet intermediaries, and internet users to report illegal content objectively, diligently, and accurately, and within the limits of their expertise. If they do not do so, they may be decertified or suspended from their notification privileges. It is suggested in the chapter that this positively contributes to accountability. The internet intermediary is accountable to the trusted flagger and to the certifying state actor for acknowledging and prioritising such notices, but similarly to internet users in weighing their interests against those of the trusted flagger. If they are incompliant, they may be fined for non-compliance with the DSA, or in some cases be held liable for not removing illegal content.

State actors are similarly accountable to the general public and internet intermediaries in selecting entities with sufficient expertise to fulfil the role of trusted flagger adequately. However, unlike the trusted flagger and the internet intermediary, it is not obvious from the DSA how a state actor would be held accountable for matters such as wrongly certifying a trusted flagger. Although this might be resolved in national administrative law, such a solution would not extend beyond national borders. Chapter V therefore argued that correction mechanisms should be instituted for state actors to be corrected in their selection of trusted flaggers. In addition, harmonising the requirements for notice-and-action

mechanisms for trusted flaggers can foster accountability for their flagging practices by providing additional transparency and ensuring that flags can be tested against basic safeguards for freedom of expression.

The role of digital services coordinators in networked platform governance has remained relatively underexplored in literature. Chapter VI has argued that they have a significant position in the protection of freedom of expression on the internet. As explained above, the enforcement structure of the DSA puts pressure on the freedom of expression. In order to ensure sufficient accountability of DSCs for the protection of this right in their enforcement of the DSA, the chapter proposed creating an accountability mechanism within the Board of Digital Services Coordinators to force DSCs to take action when so required. Additionally, it suggested that adequate interpretation of national laws could be offered by introducing a horizontal preliminary ruling procedure in cases where definitions of illegal content are conflicting or unclear. By doing so, DSCs can be held accountable by other DSCs for the correct interpretation of national laws.

Finally, Chapter VII identifies a lack of accountability in situations where regulators informally pressure internet intermediaries to change their content moderation practices. This accountability vacuum can be observed, for example, when state actors partake in drafting self-regulatory or co-regulatory instruments, such as Codes of Conduct under the DSA. Regardless of the desirability of such interventions, it is discussed that the opacity of informal pressures creates friction with requirements of article 10(2) ECHR. In addition, there is no instrument to hold state actors accountable for their input in such interventions. To fill the accountability vacuum, the chapters proposed that state actors should be transparent about their input in co-regulatory affairs, which would allow for better scrutiny of their compliance with the requirements of the Convention and the Charter.

Hence, this dissertation has demonstrated that, in complex governance networks, accountability can be a contentious element. When many actors are involved, it is not always clear who can be held accountable for the outcome of the governance network. This is also pertinent in the DSA's networked platform governance. The various chapters have shown that it is not always obvious how the conduct of various actors affects the balancing of freedom of expression in content moderation. The suggestions highlighted in this subsection are therefore either to create mechanisms for accountability within the framework of the DSA, or to foster transparency about how interactions of actors alter the weight given to the right to

freedom of expression in content moderation, which subsequently enables them being held accountable for this role.

Discussion and final remarks

This dissertation has set out to map how the DSA and the actors it formalises affect the balancing of freedom of expression in content moderation. Its contribution has been twofold. First, it has identified and characterised risks to the freedom of expression of internet users stemming from the DSA's procedural take on content moderation regarding the involvement of internet intermediaries, users, trusted flaggers, DSCs, and other state actors. Second, it has suggested interpretations of the DSA that could mitigate or help to adequately balance those risks with other interests, and increase accountability over where the balance of freedom of expression and such other interests is struck. As such, it contributes to a policy dialogue on the further interpretation and enforcement of the DSA.

Meanwhile, the enforcement of the DSA is still very much in its infant's shoes. The first cases at the national level are starting to appear,[14] and VLOPSEs are having their first battles with the Commission regarding their designation[15] or supervisory fees.[16] It seems only a matter of time before more DSA cases will be litigated.[17] Some of the provisions in the DSA are very open-ended and can create a proverbial tug-of-war between litigants and internet intermediaries on whether the DSA is to be explained pro-platform or pro-user, rightsholder, or general public. In addition to this, some national digital services coordinators have already started their first investigations. As discussed in Chapter VI, particularly the Irish DSC has been active in requesting information from predominantly social media platforms, in its own competence or on behalf of complaints submitted by Dutch and German NGOs.[18] The Commission is also preparing its first decisions on compliance with the DSA. It has cast a wide net by recently making over 50 requests for information with VLOPSEs on several aspects of their service that fall under the scope of the DSA. However, commentators

[14] See e.g. Rechtbank Amsterdam, 5 July 2024, *Mekic v X,* ECLI:NL:RBAMS:2024:3980; Landesgericht Berlin, 6 February 2025, *DRI v X,* 41 O 140/25, upcoming: Rechtbank Amsterdam, 2 October 2025, *Bits of Freedom v Meta.*

[15] E.g. Court of Justice of the European Union (General Court), 3 September 2025, *Zalando v Commission,* T-348/23; Court of Justice of the European Union (Order), 27 March 2024, *European Commission v Amazon,* C-639/23 P(R).

[16] Court of Justice of the European Union (General Court), 10 September 2025, *Meta Platforms Ireland v European Commission,* T-55/24; Court of Justice of the European union (General Court), 10 September 2025, *TikTok Technology v European Commission,* T-58/24.

[17] See generally: Paddy Leerssen and others, 'Pathways to Private Enforcement of the Digital Services Act (DSA)' (Institute for Information Law 2025).

[18] For an overview, see: https://dscdb.edri.org/EuropeanCommission

have noted that generally, enforcement efforts have been slow and opaque, and impactful decisions are still expected.[19]

The uncertain development of enforcement efforts means that, for the most part, it remains unclear how the risks addressed in this dissertation are weighed in the enforcement stages. Because the DSA is content agnostic, it can be used in a manner that is respecting and conducive to fundamental rights on the internet. At the same time, it can also be interpreted in a way that could harm the freedom of expression of internet users. The DSA can be seen as J.K. Rowling's Room of Requirement:[20] everyone will be able to give their own interpretation in its content-agnosticism to serve their own needs. The contributions made in this dissertation are a starting point for guiding an interpretation of the DSA that adequately balances freedom of expression against the interests of involved stakeholders.

To bring the discussion and enforcement of this instrument further, multidisciplinary research is paramount. Such multidisciplinary insights are required in DSA enforcement as well: the usefulness of an interpretation-heavy instrument such as the DSA is fully dependent on how well the people applying it understand the services they are applying it to. The transparency instruments in the DSA have the potential to provide unprecedented insights into the harms caused by content moderation online, which could guide our understanding of what regulatory steps are required to those harms. Both qualitative and quantitative research is necessary to paint the full picture. However, as noted in the limitation section in Chapter I, legal research is crucial to make sense of this instrument. More insights from legal scholarship are necessary to guide enforcement through the uncharted waters of overlapping instruments in the digital acquis. Equally necessary is further research to guide the interpretation of open- or relatively undefined norms in the DSA, such as 'systemic risk' and 'illegal content', as well as the development and operationalisation of broader notions such as accountability and transparency in the specific context of regulating digital services.

The broader message of this dissertation is that honesty is needed about what the DSA can and cannot do. Most policy decisions regarding the regulation of online services will have effects on the freedom of expression. These policy interventions

[19] See e.g. Matteo Fabbri, 'The Role of Requests for Information in Governing Digital Platforms Under the Digital Services Act: The Case of X' (2025) 6 Journalism and Media 41.

[20] J.K. Rowling, 'Harry Potter and the Order of the Phoenix', 2003: the Room of Requirement appears whenever someone needs it, and provides whatever the person in need requires.

may be required to pursue legitimate aims. However, it is necessary to be open about why aspects of freedom of expression can be sacrificed to serve that intervention, and how the sacrifice is mitigated as much as possible. This means also acknowledging the DSA's limitations, and accepting that it cannot solve some societal problems. This is especially pertinent since each supervisory authority will display its own policy goals onto the DSA's agnosticism. Overly broad interpretations of the DSA do not only risk making it an unpopular regulatory instrument; they risk throwing the baby out with the bathwater by jeopardising a free online environment in pursuit of policy aims.

Bibliography

Regulation

European Convention on Human Rights, 213 UNTS 222, 4 November 1950.

Charter of Fundamental Rights of the European Union, *OJ C 326*, 26 October 2012, p. 391-407.

Treaty on the European Union, *OJ C* 326, 26 October 2012, p. 13-390.

Treaty on the Functioning of the European Union, *OJ C 202*, 7 June 2016, p. 1-388.

Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, *OJ L 119*, 4 May 2016, p. 1-88.

Regulation 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, *OJ L 172*, 29 April 2021, p. 79-101.

Regulation 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, *OJ L 277*, 27 October 2022, p. 1-102.

Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, *OJ L 2024/1689*, 12 July 2024, p. 1-44

Directive 98/48 of the European Parliament and of the Council of 20 July 1998 amending Directive 98/34/EC laying down a procedure for the provision of information in the field of technical standards and regulations, *OJ L 217*, 5 August 1998, p. 18-26.

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 July 2000, p 1-16.

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, O*J L 167*, 22.6.2001, p. 10-19.

Directive 2011/93 of the European Parliament and of the Council of 13 December 2011 on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision 2004/68/JHA, *OJ L 335*, 17.12.2011, p. 1-14.

Directive 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society Services, *OJ L 241*, 17 September 2015, p. 1-15.

Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities, *OJ L 303*, 28 November 2018, p. 69–92.

Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, *OJ L 130*, 17 May 2019, p. 92-125.

Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises, notified under document number C(2003)1422, *OJ L 124*, 20 May 2003, p. 36-41.

Council Framework Decision 2008/913/JHA of 28 November 2008 on certain forms and expressions of racism and xenophobia by means of criminal law

Commission Recommendation (EU) 2018/334 of 1 March 2018 on measures to effectively tackle illegal content online, C/2018/1177, *OJ L 63*, 6 March 2018, p. 50-61.

Delegated Regulation supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines, C(2023) 6807 final.

Commission Delegated Regulation (EU) of 1 July 2025 supplementing Regulation 2022/2065 of the European Parliament and of the Council by laying down the

technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data with vetted researchers, C(2025) 4340 final.

Case law

Council of Europe case law

European Court of Human Rights, 7 December 1976, *Handyside v the United Kingdom*, 5493/72.

European Court of Human Rights, 26 April 1979, *The Sunday Times v the United Kingdom*, 6538/74.

European Court of Human Rights, 8 July 1986, *Lingens v Austria*, 9815/82.

European Court of Human Rights, 30 January 1998, *United Communist Party of Turkey and Others v Turkey*, 19392/92

European Court of Human Rights, 23 September 1998, *Lehideux & Isorni v France*, 24662/94.

European Court of Human Rights, 18 February 1999, *Matthews v United Kingdom*, 24833/94.

European Court of Human Rights, 13 February 2003, *Cetin et al v Turkey*, 40153/98 and 40160/98.

European Court of Human Rights, 24 September 2003, *Appleby et al v the United Kingdom*, 44306/98.

European Court of Human Rights, 30 June 2005, *Bosphorus v Ireland*, 45036/98.

European Court of Human Rights, 23 October 2008, *Khuzhin et al v Russia*, 13470/02.

European Court of Human Rights, 4 December 2008, *S. and Marper v the United Kingdom*, 30562/04 & 30566/04.

European Court of Human Rights, 20 January 2009, *PO Kokkelvisserij v the Netherlands*, 13645/05.

European Court of Human Rights, 16 July 2009, *Féret v Belgium*, 15615/07.

European Court of Human Rights, 16 July 2009, *Willem v France*, 10883/05.

European Court of Human Rights, 14 September 2010, *Dink v Turkey*, 2668/07.

European Court of Human Rights, 25 January 2012, *Altug Taner Akam v Turkey*, 27520/07.

European Court of Human Rights, 9 February 2012, *Vejdeland et al v Sweden,* 1813/07.

European Court of Human Rights, 2 October 2012, *Rujak v Croatia,* 57942/10.

European Court of Human Rights, 18 December 2012, *Ahmet Yilderim v Turkey*, 3111/10.

European Court of Human Rights, 19 February 2013, *Neij and Sunde Kolmisoppi v Sweden,* 40397/12.

European Court of Human Rights, 2 April 2013, *Animal Defenders International v the United Kingdom*, 48876/08.

European Court of Human Rights, 16 January 2014, *Tierbefreier eV v Germany*, 45192/09.

European Court of Human Rights, 16 June 2015, *Delfi v Estonia,* 64569/09.

European Court of Human Rights, 15 October 2015, *Perincek v Switzerland,* 27510/08.

European Court of Human Rights, 2 February 2016, *MTE and Index.hu v Hungary*, 22947/13.

European Court of Human Rights, 8 November 2016, *Magyar Helsinki Bizottsag v Hungary,* 18030/11.

European Court of Human Rights, 21 March 2017, *Janssen Cilag S.A.S. v France,* 33931/12.

European Court of Human Rights, 27 June 2017, *Belkacem v Belgium*, 34367/14.

European Court of Human Rights, 27 June 2017, *Stakunnan Markkinapörssi Oy and Satamedia Oy v Finland,* 931/13.

European Court of Human Rights, 21 November 2019, *Ilias and Ahmed v Hungary*, 47287/15.

European Court of Human Rights, 20 January 2020, *Magyar Ketfarku Kutya Part v Hungary,* 201/17.

European Court of Human Rights, 22 December 2020, *Selahattin Demirtas v Turkey*, 14305/17.

European Court of Human Rights, 18 May 2021, *OOO Informatsionnoye Agentstvo Tambov-Inform v Russia*, 43351/12.

European Court of Human Rights, 15 September 2021, *Melike v Turkey*, 35786/19.

European Court of Human Rights, 25 February 2022, *Biancardi v Italy*, 77419/16.

European Court of Human Rights, 5 April 2022, *NIT v Moldova*, 28470/12.

European Court of Human Rights, 15 May 2023, *Sanchez v France*, 45581/15.

European Court of Human Rights, 7 January 2025, *Patrascu v Romania*, 1847/21.

European Court of Human Rights, 8 July 2025, *Google v Russia*, 37027/22.

European Union case law

Court of Justice of the European Union, 16 December 1976, *Rewe v Landwirtschaftskammer für das Saarland*, C-33/76.

Court of Justice of the European Union, judgment of 9 March 2010, *European Commission v Federal Republic of Germany*, C-518/07.

Court of Justice of the European Union, 23 March 2010, *Google France v Louis Vuitton*, C-236/08 & C-238/08.

Court of Justice of the European Union, 12 July 2011, *L'Oreal v eBay*, C-324/09.

Court of Justice of the European Union, 24 November 2011, *Scarlet Extended v SABAM*, C-70/10.

Court of Justice of the European Union, 16 February 2012, *SABAM v Netlog*, C-360/10.

Court of Justice of the European Union, judgment of 16 October 2012, *European Commission v Republic of Austria*, C-614/10.

Court of Justice of the European Union, 22 December, *DEB v Germany*, C-279/09.

Court of Justice of the European Union, 27 March 2014, *UPC Telekabel Wien GmbH v. Constantin Film Verleih GmbH and Wega Filmproduktionsgesellschaft mbH*, C-314/12.

Court of Justice of the European Union, 8 April 2014, *Digital Rights Ireland*, C-293/12 & C-594/12.

Court of Justice of the European Union, 13 May 2014, *Google Spain v Mario Gonzalez*, C-131/32.

Court of Justice of the European Union, 11 September 2014, *Sotiris Papasavvas v O Fileleftheros Dimosia Etaireia,* C-291/13.

Court of Justice of the European Union, 18 December 2014, Opinion 2/13.

Court of Justice of the European Union, 4 May 2016, *Philip Morris,* C-547/14.

Court of Justice of the European Union, 15 September 2016, *Tobias McFadden v Sony Music Entertainment*, C-484/14.

Court of Justice of the European Union, 26 april 2017, *Stichting Brein v Filmspeler,* C-527/15.

Court of Justice of the European Union, 14 June 2017, *Stichting Brein v Ziggo & XS4All*, C-610/15.

Court of Justice of the European Union, judgment of 13 June 2018, *Commission v Poland,* C-530/16.

Court of Justice of the European Union, judgment of 20 June 2018, *Cesky Drahy v Commission,* T-325/16.

Court of Justice of the European Union, 3 October 2019, *Eva Glawischnig-Piesczek v Facebook,* C-18/18.

Court of Justice of the European Union, 6 October 2020, *La Quadrature du Net,* C-511/18, C-512/18 & C-520/18.

Court of Justice of the European Union, 15 June 2021, *Facebook v Gegevensbeschermingsautoriteit,* C-645/19.

Court of Justice of the European Union, 22 June 2021, *Peterson v Google/Elsevier v Cyando,* C-682/18 & C-683/18.

Court of Justice of the European Union, 26 April 2022, *Commission v Poland,* C-401/19.

Court of Justice of the European Union, 4 July 2023, *Meta Platforms Inc v Bundeskartellamt*, C-252/21.

Court of Justice of the European Union (Order), 27 March 2024, *European Commission v Amazon,* C-639/23 P(R).

Court of Justice of the European Union (General Court), 3 September 2025, *Zalando v Commission,* T-348/23.

Court of Justice of the European Union (General Court), 10 September 2025, *Meta Platforms Ireland v European Commission*, T-55/24.

Court of Justice of the European union (General Court), 10 September 2025, *TikTok Technology v European Commission*, T-58/24.

National case law

France
Conseil Constitutionnel, 18 June 2020, 2020-801 DC.

Germany
Bundesgerichtshof, 12 March 2004, *Internetversteigerung I*, I ZR 304/01.

Bundesverfassungsgericht, 22 May 2019, *Der III. Weg*, 1 BvQ 42/19.

Bundesgerichtshof, 29 July 2021, *Facebook*, III ZR 179/20.

Landesgericht Berlin, 6 February 2025, *DRI v X*, 41 O 140/25.

Kammergericht Berlin, 25 August 2025, 10 W 70/25.

Israel
High Court of Justice, 12 April 2021, *Adalah Legal Center for Arab Minority Rights in Israel v State Attorney's Office, Cyber Department*, HCJ 7846/19.

The Netherlands
Hoge Raad, 27 January 2023, *Brein v NSE*, ECLI:NL:HR:2023:94.

Rechtbank Amsterdam, 5 July 2024, *Mekic v X*, ECLI:NL:RBAMS:2024:3980.

Rechtbank Amsterdam, 2 October 2025, *Bits of Freedom v Meta.*

Sweden
Swedish District Court, 19 May 2015, *.SE.*

United States
California Court of Appeals, June 23 1958, *People v Smith*, Civ. A. No. 3792.

Supreme Court, 19 June 2017, *Packingham v North Carolina*, 137 S. Ct. 1730.

Court of Appeals 5th Circuit, 13 January 2022, *Biden v Missouri*, 595 U.S. __ (2022).

Supreme Court, 26 June 2024, *Murthy v Missouri*, 601 U.S. __(2024).

Other
Meta Oversight Board, 22 November 2022, *UK Drill Music.*

Meta Oversight Board, 31 August 2023, *Holocaust Denial.*

Literature

Adams M, Ballin EH and Meuwese A, 'The Ideal and the Real in the Realm of Constitutionalism and the Rule of Law: An Introduction' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017)

Agarwal V and Sastry N, '"Way Back Then": A Data-Driven View of 25+ Years of Web Evolution', *Proceedings of the ACM Web Conference 2022* (ACM 2022).

Alemanno A and Sibony A-L, *Nudge and the Law : A European Perspective* (1st edn, Hart Publishing 2015)

Alemanno A and Veraldi J, *Musk, Power, and the EU: Can EU Law Tackle the Challenges of Unchecked Plutocracy?* (Verfassungsbooks 2025)

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

—, *Hate Speech and the European Court of Human Rights* (Routledge 2025)

Ananny M and Crawford K, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973

Andersson H, *Dawn Raids Under Challenge: Due Process Aspects on the European Commission's Dawn Raid Practices* (Bloomsbury Publishing 2024)

Angelopoulos C and others, 'Study of Fundamental Rights Limitations for Online Enforcement through Self-Regulation' (Institute for Information Law 2016)

Angelopoulos C and and Smet S, 'Notice-and-Fair-Balance: How to Reach a Compromise between Fundamental Rights in European Intermediary Liability' (2016) 8 Journal of Media Law 266

Annabell T, Aade L and Goanta C, '(Un)Disclosed Brand Partnerships: How Platform Policies and Interfaces Shape Commercial Content for Influencers' (2024) 13 Internet Policy Review

Appelman N and Leerssen P, 'On Trusted Flaggers' (2022) 24 Yale Journal of Law & Technology 452

Are C, 'Flagging as a Silencing Tool: Exploring the Relationship between de-Platforming of Sex and Online Abuse on Instagram and TikTok' [2024] New Media & Society 14614448241228544

Balkin JM, 'Old-School/New-School Speech Regulation' (2014) 127 Harvard Law Review 2296

—, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 University of California, Davis Law Review 1149

—, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011

Bambauer D, 'Orwell's Armchair' (2012) 79 University of Chicago Law Review 863

—, 'Against Jawboning' (2015) 100 Minnesota Law Review 51

Barata J, 'The Future of Free Speech: Old Threats and New Challenges' in Eric Heinze and others (eds), *Oxford Handbook of Hate Speech* (Oxford University Press forthcoming)

Barendt E, 'Freedom of Expression' in Michel Rosenfeld and Andras Sajo, *The Oxford Handbook of Comparative Constitutional law* (Oxford University Press 2012)

Barendt EM, *Freedom of Speech* (2nd ed, Oxford University Press 2007)

Benedek W and Kettemann MC, 'Freedom of Expression and the Internet' (Council of Europe 2013)

Benkler Y, *The Wealth of Networks: How Social Production Transforms Markets and Freedom Contract : Freedom in the Commons* (Yale University Press 2006)

—, 'A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate' (2011) 46 Harvard Civil Rights Civil Liberties Law Review 311

Berbec A, 'To What Extent Can Online Service Providers Adopt Voluntary Content Moderation Measures without Losing Their Immunity Shields? A Comparative Analysis of Online Service Providers' Liabilities in the European Union and the United States' (2024) 15 JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law

Bertaglia T, Goanta C and Iamnitchi A, 'The Monetisation of Toxicity: Analysing YouTube Content Creators and Controversy-Driven Engagement', *4th International Workshop on OPEN CHALLENGES IN ONLINE SOCIAL NETWORKS* (ACM 2024)

Bietti E, 'A Genealogy of Digital Platform Regulation' (2023) 7 Georgetown Law and Technology Review 1

Birnhack M D and Elkin-Koren N, 'The Invisible Handshake: The Reemergence of the State in the Digital Environment' (2003) 8 Virginia Journal of Law & Technology 1

Bisarya S and Bulmer WE, 'Rule of Law, Democracy and Human Rights: The Paramountcy of Moderation' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017)

Bloch-Wehba H, 'Global Platform Governance: Private Power in the Shadow of the State' (2019) 72 SMU Law Review 28

Bogason P and Musso JA, 'The Democratic Prospects of Network Governance' (2006) 36 The American Review of Public Administration 3

Botero Arcila B and Griffin R, 'Social Media Platforms and Challenges for Democracy, Rule of Law and Fundamental Rights' (European Parliament 2023) PE 743.400

Bovens M, 'Analysing and Assessing Accountability: A Conceptual Framework' (2007) 13 European Law Journal 447

Boyd DM and Ellison NB, 'Social Network Sites: Definition, History, and Scholarship' (2007) 13 Journal of Computer-Mediated Communication 210

Bozarth L and others, 'Wisdom of Two Crowds: Misinformation Moderation on Reddit and How to Improve This Process---A Case Study of COVID-19' (2023) 7 Proceedings of the ACM on Human-Computer Interaction 1

Bradford A, 'The Brussels Effect' (2012) 107 Northwestern University Law Review 1

Bridy A, 'The Price of Closing the "Value Gap": How the Music Industry Hacked EU Copyright Reform' (2020) 22 Vanderbilt Journal of Entertainment & Technology Law 323

Brito Bastos F and Pałka P, 'Is Centralised General Data Protection Regulation Enforcement a Constitutional Necessity?' (2023) 19 European Constitutional Law Review 487

Broughton Micova S and Calef A, 'Elements for Effective Systemic Risk Assessment Under the DSA' (Centre on Regulation in Europe 2023)

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

Bucher T, 'Want to Be on the Top? Algorithmic Power and the Threat of Invisibility on Facebook' (2012) 14 New Media & Society 1164

Bucher T and Helmond A, 'The Affordances of Social Media Platforms' in Jean Burgess, Alice Marwick and Thomas Poell, *The SAGE Handbook of Social Media* (SAGE Publications Ltd 2018)

Buiten MC, De Streel A and Peitz AM, 'Rethinking Liability Rules for Online Hosting Platforms' (2020) 28 International Journal of Law and Information Technology 139

Caplan R, 'Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches' (Data & Society Research Institute 2018)

—, 'Networked Platform Governance: The Construction of the Democratic Platform' (2023) 17 International Journal of Communication 3451

Caplan R and Gillespie T, 'Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy' (2020) 6 Social Media + Society 205630512093663

Casarosa F, 'When the Algorithm Is Not Fully Reliable: The Collaboration between Technology and Humans in the Fight against Hate Speech' in Hans-W Micklitz and others (eds), *Constitutional Challenges in the Algorithmic Society* (1st edn, Cambridge University Press 2021)

Castets-Renardt C, 'Algorithmic Content Moderation on Social Media in EU Law: Illusion of Perfect Enforcement' (2020) 2020 University of Illinois Journal of Law Technology & Policy 283

Cauffman C and Goanta C, 'A New Order: The Digital Services Act and Consumer Protection' (2021) 12 European Journal of Risk Regulation 758

Celeste E, 'Terms of Service and Bills of Rights: New Mechanisms of Constitutionalisation in the Social Media Environment?' (2019) 33 International Review of Law, Computers & Technology 122

Cerf V, Ryan P (I) and Senges M, 'Internet Governance Is Our Shared Responsibility' (2014) 10 I/S: A Journal of Law and Policy for the Information Society 1

Chang B, 'From Internet Referral Units to International Agreements; Censorship of the Internet by the UK and EU' (2018) 49 Columbia Human Rights Law Review 113

Chang H-CH, Richardson A and Ferrara E, '#JusticeforGeorgeFloyd: How Instagram Facilitated the 2020 Black Lives Matter Protests' (2022) 17 PLOS ONE e0277864

Chee FY, 'EU's Breton Wants Google, YouTube to Ban War Propaganda Accounts' *Reuters* (27 February 2022) <https://www.reuters.com/technology/eus-breton-hold-video-call-with-google-youtube-ceos-ukraine-fake-news-2022-02-27/> accessed 28 November 2022

Chin YC and Changfeng C, 'Internet Governance: Exploration of Power Relationship' (2017) <https://papers.ssrn.com/abstract=3107239> accessed 21 January 2025

Citron DK, 'Extremist Speech, Compelled Conformity, and Censorship Creep' (2018) 93 Notre Dame Law Review 1035

Clark L, 'Facebook and Twitter Must Tackle Hate Speech or Face New Laws' *Wired UK* <https://www.wired.co.uk/article/us-tech-giants-must-tackle-hate-speech-or-face-legal-action> accessed 1 November 2022

Cleynenbreugel PV and Mattioli P, 'Digital Services Coordinators and Other Competent Authorities in the Digital Services Act: Streamlined Enforcement Coordination Lost?' [2023] European Law Blog <https://europeanlawblog.pubpub.org/pub/digital-services-coordinators-and-other-competent-authorities-in-the-digital-services-act-streamlined-enforcement-coordination-lost> accessed 19 August 2024

Coche E, 'Privatised Enforcement and the Right to Freedom of Expression in a World Confronted with Terrorism Propaganda Online' (2018) 7 Internet Policy Review

'Commission opens formal proceedings against TikTok under DSA' (*European Commission – European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487> accessed 18 December 2024

'Commission Opens Formal Proceedings under DSA' (*European Commission – European Commission*) <https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373> accessed 26 May 2025

'Conference's notes – Getting Ready for the DSA: The Role of Trusted Flaggers for a Better Moderation of Platforms' (*Régulation du Numérique*, 8 February 2024) <https://regulation-tech.cnam.fr/summary-arcom-event-on-january-26th-2024-getting-ready-for-the-dsa-the-role-of-trusted-flaggers-for-a-better-moderation-of-platforms/> accessed 14 March 2024

'Content Moderation in a New Era for AI and Automation | Oversight Board' (17 September 2024) <https://www.oversightboard.com/news/content-moderation-in-a-new-era-for-ai-and-automation/> accessed 27 February 2025

Cornlis M, Auler L and Kirsch S, 'Vollzug des Digital Services Act in Deutschland – Implementierung einer verbraucherorientierten Aufsichtsbehördenstruktur' (Mainzer Medieninstitut 2022)

Costello RÁ, 'Faux Ami? Interrogating the Normative Coherence of "Digital Constitutionalism"' (2023) 12 Global Constitutionalism 326

Cox AL and others, 'Design Frictions for Mindful Interactions: The Case for Microboundaries', *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Association for Computing Machinery 2016)

Crawford K and Gillespie T, 'What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint' (2016) 18 New Media & Society 410

Dang S, 'Twitter Dissolves Trust and Safety Council' *Reuters* (13 December 2022) <https://www.reuters.com/technology/twitter-dissolves-trust-safety-council-2022-12-13/> accessed 18 July 2025

Datta A, 'US Big Tech Backtracks on EU Fact-Checking Commitments' (*Euractiv*, 22 January 2025) <https://www.euractiv.com/section/tech/news/us-big-tech-backtracks-on-eu-fact-checking-commitments/> accessed 18 July 2025

—, 'Virkkunen: All Digital Rules to Be Reviewed in Simplification Push' *Euractiv* (27 March 2025) <https://www.euractiv.com/section/tech/news/virkkunen-all-digital-rules-to-be-reviewed-in-simplification-push/> accessed 7 April 2025

de Abreu Duarte F, De Gregorio G and Golia A, 'Perspectives on Digital Constitutionalism' in O Kanevskaia, P Palka and B Brozek (eds), *Handbook on Law & Technology* (Edward Elgar 2023)

de Carvalho MC, 'It Will Be What We Want It to Be: Sociotechnical and Contested Systemic Risk at the Core of the EU's Regulation of Platforms' AI Systems' [2025]

Journal of Intellectual Property, Information Technology and Electronic Commerce Law 35

De Gregorio G, 'Expressions on Platforms': (2018) 2 European Competition and Regulatory Law Review 203

—, 'The Rise of Digital Constitutionalism in the European Union' (2021) 19 International Journal of Constitutional Law 41

—, 'Digital Constitutionalism across the Atlantic' [2022] Global Constitutionalism 1

—, *Digital Constitutionalism in Europe: Reframing Rights and Powers in the Algorithmic Society* (1st edn, Cambridge University Press 2022)

De Gregorio G and Demková S, 'The Constitutional Right to an Effective Remedy in the Digital Age: A Perspective from Europe' in Charlotte Van Oirsouw and others (eds), *European Yearbook of Constitutional Law 2023*, vol 5 (TMC Asser Press 2024)

De Gregorio G and Dunn P, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473

De Gregorio G and Goanta C, 'The Influencer Republic: Monetizing Political Speech on Social Media' [2022] German Law Journal 204

de Laat PB, 'Coercion or Empowerment? Moderation of Content in Wikipedia as "Essentially Contested" Bureaucratic Rules' (2012) 14 Ethics and Information Technology 123

De Streel A, 'Online Platforms' Moderation of Illegal Content Online' (European Parliament 2020) PE 652.718

De Vreese C, Ohme J and Seiling L, 'Will Europe Sacrifice the Digital Services Act in Negotiations with Trump? | TechPolicy.Press' *Tech Policy Press* (25 March 2025) <https://techpolicy.press/will-europe-sacrifice-the-digital-services-act-in-negotiations-with-trump> accessed 7 August 2025

Decuypere M, 'The Topologies of Data Practices: A Methodological Introduction' (2021) 10 Journal of New Approaches in Educational Research 67

Del Campo A, Zara N and Álvarez-Ugarte R, 'Are Risks the New Rights? The Perils of Risk-Based Approaches to Speech Regulation' (2025) 16 Journal of Intellectual Property, Information Technology and Electronic Commerce

Demková S and De Gregorio G, 'The Looming Enforcement Crisis in European Digital Policy' (*Verfassungsblog*, 10 February 2025) <https://verfassungsblog.de/the-looming-enforcement-crisis-ai-dsa-eu/> accessed 28 February 2025

Democracy C for, 'Far Right News Sources on Facebook More Engaging' (*Medium*, 4 March 2021) <https://medium.com/cybersecurity-for-democracy/far-right-news-sources-on-facebook-more-engaging-e04a01efae90> accessed 12 April 2025

Dergacheva D and others, 'One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database' (Center for Media, Communication and Information Research 2023) <https://platform-governance.org/2023/one-day-in-content-moderation-by-social-media-platforms-in-the-eu/> accessed 14 December 2023

Dijck J van, Poell T and Waal M de, *The Platform Society: Public Values in a Connective World* (Oxford University Press 2018)

Douek E, 'The Limits of International Law in Content Moderation' (2020) 6 U.C. Irvine Journal of International, Transnational and Comparative Law 37

—, 'Content Moderation as Systems Thinking' (2022) 136 Harvard Law Review 526

—, 'The Meta Oversight Board and the Empty Promise of Legitimacy' (2024) 37 Harvard Journal of Law & Technology

Draper NA, 'Distributed Intervention: Networked Content Moderation in Anonymous Mobile Spaces' (2019) 19 Feminist Media Studies 667

Drolsbach C and Pröllochs N, 'Content Moderation on Social Media in the EU: Insights From the DSA Transparency Database' (arXiv, 7 December 2023) <http://arxiv.org/abs/2312.04431> accessed 14 December 2023

Drolsbach CP, Solovev K and Pröllochs N, 'Community Notes Increase Trust in Fact-Checking on Social Media' (2024) 3 PNAS Nexus pgae217

Duguay S and Gold-Apel H, 'Stumbling Blocks and Alternative Paths: Reconsidering the Walkthrough Method for Analyzing Apps' (2023) 9 Social Media + Society 20563051231158822

Dunn P, 'Online Hate Speech and Intermediary Liability in the Age of Algorithmic Content Moderation' (University of Bologna 2024)

Dvoskin B, 'Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance' (2022) 67 Villanova Law Review 447

Echikson W and Knodt O, 'Germany's NetzDG: A Key Test for Combatting Online Hate' (Centre for European Policy Studies 2018) 2018/09

Edelson L, Graef I and Lancieri F, 'Access to Data and Algorithms: For an Effective DMA and DSA Implementation' (Centre on Regulation in Europe 2023)

Eder N, 'Making Systemic Risk Assessments Work: How the DSA Creates a Virtuous Loop to Address the Societal Harms of Content Moderation' [2024] German Law Journal 1

Edwards L, *Law, Policy, and the Internet* (Hart publishing 2019)

Edwards L and Veale M, 'Slave to the Algorithm? Why a Right to an Explanation Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18

Eghbariah R and Metwally A, 'Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service' (2021) 23 Yale Journal of Law & Technology 542

Eifert M (ed), *Netzwerkdurchsetzungsgesetz in Der Bewährung: Juristische Evaluation Und Optimierungspotenzial* (1. Auflage, Nomos 2020)

Elkin-Koren N, Gregorio GD and Perer M, 'Social Media as Contractual Networks: A Bottom Up Check on Content Moderation' (2022) 107 Iowa Law Review 987

Emaus JM, 'European Union Co- and Self-Regulation and the Protection of Fundamental Rights' in Madeleine De Cock Buning and Linda Senden (eds), *Private Regulation and Enforcement in the EU* (Hart 2020)

Endicott T, 'The Coxford Lecture: Arbitrariness' (2014) 27 Canadian Journal of Law & Jurisprudence 49

Engel A, Groussot X and Petursson GT (eds), *New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights*, vol 13 (Springer Nature Switzerland 2025)

'EU Should Simplify Its AI Act and Digital Rulebook: Danish Minister' (*euronews*, 15:34:17 +02:00) <https://www.euronews.com/next/2025/07/16/eu-should-simplify-its-ai-act-and-digital-rulebook-says-danish-minister> accessed 24 July 2025

'Europe's Enforcement Paralysis' (Irish Council for Civil Liberties 2021) <https://www.iccl.ie/wp-content/uploads/2021/09/Europes-enforcement-paralysis-2021-ICCL-report-on-GDPR-enforcement.pdf> accessed 24 October 2024

Fabbri M, 'The Role of Requests for Information in Governing Digital Platforms Under the Digital Services Act: The Case of X' (2025) 6 Journalism and Media 41

Farrand B and Carrapico H, 'Networked Governance and the Regulation of Expression on the Internet: The Blurring of the Role of Public and Private Actors as Content Regulators' (2013) 10 Journal of Information Technology & Politics 357

Feezell JT and others, 'Who Gets Flagged? An Experiment on Censorship and Bias in Social Media Reporting' (2023) 56 PS: Political Science & Politics 222

FitzGerald M, 'Not Hollowed by a Delphic Frenzy: European Intermediary Liability from the Perspective of a Bad Man: A Response to Martin Husovec et Al.' (2025) 32 Maastricht Journal of European and Comparative Law 47

Flyverbom M, 'Transparency: Mediation and the Management of Visibilities' (2016) 10 International Journal of Communication 110

Freeman J, 'The Private Role in the Public Governance' (2000) 75 New York University Law Review 543

Frosio G and Geiger C, 'Taking Fundamental Rights Seriously in the Digital Services Act's Platform Liability Regime' (2023) 29 European Law Journal 31

Frosio G and Husovec M, 'Accountability and Responsibility of Online Intermediaries' in Giancarlo Frosio (ed), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020)

Frosio GF, 'From Horizontal to Vertical: An Intermediary Liability Earthquake in Europe' (2017) 12 Journal of Intellectual Property Law & Practice 565

Fung A, Graham M and Weil D, *Full Disclosure: The Perils and Promise of Transparency* (Cambridge University Press 2007)

'Fury over Facebook "Napalm Girl" Censorship' *BBC News* (9 September 2016) <https://www.bbc.com/news/technology-37318031> accessed 27 August 2025

Gardbaum S, 'The Structure of a Free Speech Right' in Adrienne Stone and Frederick Schauer (eds), *The Oxford Handbook of Freedom of Speech* (Oxford University Press 2021)

'GDPR in Practice: Experiences of Data Protection Authorities' (European Union Agency for Fundamental Rights 2024)

Geiger C, Frosio G and Izyumenko E, 'Intermediary Liability and Fundamental Rights' in Giancarlo Frosio (ed), *Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020)

Geiger C and Jütte BJ, 'Platform Liability under Article 17 of the Copyright in the Digital Single Market Directive, Automated Filtering and Fundamental Rights: An Impossible Match' (2021) 70 GRUR International 517

Gellert R and Wolters P, 'The Revision of the European Framework for the Liability and Responsibilities of Hosting Service Providers' (Ministry of Economic Affairs and Climate Policy 2021) 202006403

Gentile G, 'Between Online and Offline Due Process: The Digital Services Act' in Annegret Engel, Xavier Groussot and Gunnar Thor Petursson (eds), *New Directions in Digitalisation: Perspectives from EU Competition Law and the Charter of Fundamental Rights* (Springer Nature Switzerland 2025)

Gentile G and Lynskey O, 'Deficient by Design? The Transnational Enforcement of the GDPR' (2022) 71 International and Comparative Law Quarterly 799

Gerards J, 'The Right to Freedom of Expression and of Information' in Janneke Gerards (ed), *Fundamental Rights: the European and International Dimension* (Cambridge University Press 2023)

Gerring J, *Case Study Research: Principles and Practices* (Second edition, Cambridge University Press 2017)

Gill K, 'Regulating Platform's Invisible Hand: Content Moderation Policies and Processes' (2021) 21 Wake Forest Journal of Business and Intellectual Property Law 171

Gill L, Redeker D and Gasser U, 'Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights' (2015) 15 Berkman Klein Center for Internet & Society Research 29

Gillespie T, 'Governance of and by Platforms' in Jean Burgess, Thomas Poell and Alice Marwick, *The SAGE Handbook of Social Media* (SAGE Publications 2017)

—, *Custodians of the Internet* (1st edn, Yale University Press 2018)

—, 'Content Moderation, AI, and the Question of Scale' (2020) 7 Big Data & Society 205395172094323

—, 'Do Not Recommend? Reduction as a Form of Content Moderation' (2022) 8 Social Media + Society 205630512211175

Goanta C, 'The New Social Media: Contracts, Consumers, and Chaos' (2023) 108 Iowa Law Review 118

—, 'How the White House Is Gaslighting the World About Europe's Digital Laws | TechPolicy.Press' (*Tech Policy Press*, 24 March 2025) <https://techpolicy.press/how-the-white-house-is-gaslighting-the-world-about-europes-digital-laws> accessed 7 April 2025

—, 'The Great Data Standoff: Researchers vs. Platforms Under the Digital Services Act' (arXiv, 2 May 2025) <http://arxiv.org/abs/2505.01122> accessed 23 June 2025

Goanta C, Bertaglia T and Iamnitchi A, 'The Case for a Legal Compliance API for the Enforcement of the EU's Digital Services Act on Social Media Platforms' 9

Goanta C and Spanakis J, 'Discussing The Legitimacy of Digital Market Surveillance' [2022] Stanford Journal of Computational Antitrust 44

Goldman E, 'Content Moderation Remedies' (2021) 28 Michigan Technology Law Review 1

Golia AJ and Teubner G, 'Societal Constitutionalism: Background, Theory, Debates' (2021) 15 ICL Journal 357

González Fuster G and others, 'The Right to Lodge a Data Protection Complaint: OK, but Then What?' (AccessNow 2022)

Gorwa R, 'The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content' (2019) 8 Internet Policy Review 1

—, 'What Is Platform Governance?' (2019) 22 Information, Communication & Society 854

—, 'Elections, Institutions, and the Regulatory Politics of Platform Governance: The Case of the German NetzDG' (2021) 45 Telecommunications Policy 102145

—, *The Politics of Platform Regulation: How Governments Shape Online Content Moderation* (Oxford University Press 2024)

Gorwa R, Binns R and Katzenbach C, 'Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance' (2020) 7 Big Data & Society 2053951719897945

Griffin R, 'From Brand Safety to Suitability: Advertisers in Platform Governance' (2023) 12 Internet Policy Review

—, 'Public and Private Power in Social Media Governance: Multistakeholderism, the Rule of Law and Democratic Accountability' (2023) 14 Transnational Legal Theory 46

—, 'The Law and Political Economy of Online Visibility: Market Justice in the Digital Services Act' (2023) 2023 Technology and Regulation 69

—, 'Codes of Conduct in the Digital Services Act: Functions, Benefits & Concerns' (2024) 2024 Technology and Regulation 167

—, 'Governing Platforms through Corporate Risk Management: The Politics of Systemic Risk in the Digital Services Act' [2025] European Law Open 1

—, 'The Politics of Risk in the Digital Services Act: A Stakeholder Mapping and Research Agenda' (2025) 5 Weizenbaum Journal of the Digital Society

Grochowski M, 'Shadow Contract Law in the Platform Economy' in Federica Casarosa and Mateusz Grochowski (eds), *Enforcing Private Regulation in the Platform Economy* (Mohr Siebeck 2025)

Hänska M and Bauchowitz S, 'Tweeting for Brexit: How Social Media Influenced the Referendum' in John Mair and others (eds) (Abramis Academic Publishing 2017)

Harrison R, Shipp J and Curtis A, 'Settling DSA-Related Disputes Outside the Courtroom: The Opportunities and Challenges Presented by Article 21 of the Digital Services Act' (2024)

Hartmann IA, 'A New Framework For Online Content Moderation' (2020) 36 Computer Law & Security Review 1

hateaid, 'HateAid Files a Complaint against TikTok' (*HateAid*, 14 February 2025) <https://hateaid.org/en/systemic-failure-hateaid-files-complaint-against-tiktok/> accessed 26 May 2025

Haupt CE, 'Platforms as Trustees: Information Fiduciaries and the Value of Analogy' (2020) 134 Harvard Law Review Forum 34

Haupt CE, 'The Horizontal Effect of Fundamental Rights' in Giovanni De Gregorio, Oreste Pollicino and Peggy Valcke (eds), *Oxford Handbook on Digital Constitutionalism* (Oxford University Press 2024)

Hayek F, 'Decline of the Rule of Law' [1953] The Freeman 518

Helberger N, 'On the Democratic Role of News Recommenders' (2019) 7 Digital Journalism 993

Heldt A, 'Reading between the Lines and the Numbers: An Analysis of the First NetzDG Reports' (2019) 8 Internet Policy Review

—, *Intensivere Drittwirkung* (Mohr Siebeck 2023)

Heldt A, 'Bypassing Classical Concepts of Censorship?' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 56

Henley J, 'Romania Regulator Calls for TikTok Suspension amid Vote Interference Fears' *The Guardian* (27 November 2024) <https://www.theguardian.com/world/2024/nov/27/romanian-regulator-tiktok-suspended-cyber-interference-election-georgescu> accessed 14 January 2025

Henrique Muniz Da Conceicao L, 'A Constitutional Reflector? Assessing Societal and Digital Constitutionalism in Meta's Oversight Board' [2024] Global Constitutionalism 1

Hert P de and Hajduk P, 'EU Cross-Regime Enforcement, Redundancy and Interdependence. Addressing Overlap of Enforcement Structures in the Digital Sphere after Meta.' (2024) 2024 Technology and Regulation 291

Hickey D and others, 'X under Musk's Leadership: Substantial Hate and No Reduction in Inauthentic Activity' (2025) 20 PLOS ONE e0313293

Hoffmann A and Gasparotti A, 'Liability for Illegal Content Online' (Centre for European Policy 2020)

Hofmann F and Raue B (eds), *Digital Services Act: Article-by-Article Commentary* (First edition, Nomos 2025)

Hofmann H, 'Composite Decision Making Procedures in EU Administrative Law' in Herwig CH Hofmann and Alexander H Türk (eds), *Legal Challenges in EU Administrative Law* (Edward Elgar Publishing 2009)

—, 'Multi-Jurisdictional Composite Procedures – The Backbone to the EU's Single Regulatory Space' (2019) <https://www.ssrn.com/abstract=3399042> accessed 24 October 2024

Holznagel D, 'Chapter II Des Vorschlags Der EU-Kommission Für Einen Digital Services Act – Versteckte Weichenstellungen Und Ausstehende Reparaturen Bei Den Regelungen Zu Privilegierung, Haftung & Herkunftslandprinzip Für Provider Und Online-Plattformen' (2021) 37 Computer und Recht 123

—, 'Ireland Cannot Do It Alone' (*Verfassungsblog*, 27 April 2021) <https://verfassungsblog.de/dsa-ireland/> accessed 23 April 2025

—, 'Nach dem EuGH-Urteil in Sachen YouTube/Cyando: Fast alles geklärt zur Host-Provider-Haftung?' (2021) 37 Computer und Recht 603

—, 'Safeguarding Adequate Rights Enforcement through the Digital Services Act (DSA)' <https://hateaid.org/wp-content/uploads/2022/05/2021-09-08-DSA-How-to-improve-enforcement-through-the-DSA.docx.pdf> accessed 26 August 2024

—, 'Follow Me to Unregulated Waters!' [2024] Verfassungsblog <https://verfassungsblog.de/follow-me-to-unregulated-waters/> accessed 23 September 2024

—, 'How to Apply the Notice and Action Requirements under Art. 16(6) Sentence 1 DSA – Which Action Actually?' (2024) 25 Computer Law Review International 172

Hou Y and others, 'Social Media Addiction: Its Impact, Mediation, and Intervention' (2019) 13 Cyberpsychology: Journal of Psychosocial Research on Cyberspace

House TW, 'Fact Sheet: President Donald J. Trump Issues Directive to Prevent the Unfair Exploitation of American Innovation' (*The White House*, 21 February 2025) <https://www.whitehouse.gov/fact-sheets/2025/02/fact-sheet-president-donald-j-trump-issues-directive-to-prevent-the-unfair-exploitation-of-american-innovation/> accessed 31 May 2025

Huppé GA, Creech H and Knoblauch D, 'The Frontiers of Networked Governance' (International Institute for Sustainable Development 2012)

Husovec M, *Injunctions against Intermediaries in the European Union: Accountable but Not Liable?* (1st edn, Cambridge University Press 2017)

—, 'The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown: Which Is Superior: And Why' (2018) 42 Columbia Journal of Law & the Arts 53

—, 'Mandatory Filtering Does Not Always Violate Freedom of Expression: Important Lessons from Poland v. Council and European Parliament' (2023) 60 Common Market Law Review 173

—, 'Rising Above Liability: The Digital Services Act as a Blueprint for the Second Generation Of Global Internet Rules' (2023) 38 Berkeley Technology Law Journal 883

—, *Principles of the Digital Services Act* (Oxford University Press 2024)

—, 'Grand Confusion after Sanchez v. France: Seven Reasons for Concern about Strasbourg Jurisprudence on Intermediaries' (2024) 31 Maastricht Journal of European and Comparative Law 385

Husovec M and Roche Laguna I, 'Digital Services Act: A Short Primer' [2022] SSRN Electronic Journal <https://www.ssrn.com/abstract=4153796> accessed 7 September 2022

Huszár F, Ktena SI, O'Brien C, Bellia L, Schlaikjer A and Hardt M, 'Algorithmic amplification of politics on Twitter' (2022) 119 PNAS e2025334119.

Ibrahim Y, 'The Breastfeeding Controversy and Facebook: Politicization of Image, Privacy and Protest' (2010) 1 International Journal of E-Politics (IJEP) 16

Imamović Š, 'A European Web of Human Rights Regimes: The EU, the ECHR and the National Legal Systems', *The Architecture of Fundamental Rights in the European Union* (Hart Publishing 2022)

'Impact Assessment of the Digital Services Act | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/library/impact-assessment-digital-services-act> accessed 18 July 2025

Irion K, 'The Independence of Media Regulatory Authorities in Europe' (European Audiovisual Observatory 2019)

Jahangir R, 'What's Behind Europe's Push to "Simplify" Tech Regulation? | TechPolicy.Press' (*Tech Policy Press*, 24 April 2025) <https://techpolicy.press/whats-behind-europes-push-to-simplify-tech-regulation> accessed 24 July 2025

Janal R, 'Haftung Und Verantwortung Im Entwurf Des Digital Services Acts' [2021] Zeitschrift für Europäisches Privatrecht 227

Jancewicz K, van Cleynenbreugel P and Bois J, 'Inconsistent Administrative Enforcement of EU Law at Member State Level: The Lisbon Treaty's Hidden Constitutional Challenge' (2024)

<https://orbi.uliege.be/bitstream/2268/314238/1/EUDAIMONIA_WP20242.pdf> accessed 17 October 2024

Jhaver S, Bruckman A and Gilbert E, 'Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit' (2019) 3 Proc. ACM Hum.-Comput. Interact. 150:1

Juneja P, Rama Subramanian D and Mitra T, 'Through the Looking Glass: Study of Transparency in Reddit's Moderation Practices' (2020) 4 Proc. ACM Hum.-Comput. Interact. 17:1

Kaesling K, 'Vertrauen als Topos der Regulierung vertrauenswürdiger Hinweisgeber im Digital Services Act' (2022) 86 UFITA 328

Kaesling K and Wolf A, 'Sustainability and Risk Management under the Digital Services Act: A Touchstone for the Interpretation of "Systemic Risks"' (2025) 74 GRUR International 119

Kangur U, Chakraborty R and Sharma R, 'Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes' (arXiv, 18 June 2024) <http://arxiv.org/abs/2406.12444> accessed 2 April 2025

Kaplan AM and Haenlein M, 'Users of the World, Unite! The Challenges and Opportunities of Social Media' (2010) 53 Business Horizons 59

Karanicolas M, 'Squaring the Circle Between Freedom of Expression and Platform Law' (2020) 20 Pittsburgh Journal of Technology Law & Policy 177

—, 'Even in a Pandemic, Sunlight Is the Best Disinfectant: COVID-19 and Global Freedom of Expression' (2021) 22 Oregon Review of International Law 1

Kaushal R and others, 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2024)

Kaye D, 'Report of the Special Rapporteur on the Promotion of the Right to Freedom of Opinion and Expression' (Human Rights Council 2016) A/HRC/32/38 <https://primarysources.brillonline.com/browse/human-rights-documents-online/promotion-and-protection-of-all-human-rights-civil-political-economic-social-and-cultural-rights-including-the-right-to-development;hrdhrd99702016149> accessed 11 November 2024

—, *Speech Police: The Global Struggle to Govern the Internet* (1st edn, Columbia Global Reports 2019)

Keller D, 'Internet Platforms: Observations on Speech, Danger, and Money' (Hoover Institution 2018) 1807

—, 'Dolphins in the Net: Internet Content Filters and the Advocate General's Glawischnig-Piesczek v. Facebook Ireland Opinion' (2019) <https://cyberlaw-stanford-edu.proxy.library.uu.nl/files/Dolphins-in-the-Net-AG-Analysis.pdf> accessed 25 November 2021

—, 'Who Do You Sue? State and Platform Hybrid Power over Online Speech' [2019] Hoover Working Group on National Security, Technology and Law

—, 'Facebook Filters, Fundamental Rights, and the CJEU's Glawischnig-Piesczek Ruling' (2020) 69 GRUR International 616

—, 'Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard' [2021] Journal of Free Speech Law 227

Kira B, 'Regulatory Intermediaries in Content Moderation' (2025) 14 Internet Policy Review

Klausa T, 'Graduating from "New-School" – Germany's Procedural Approach to Regulating Online Discourse' [2022] Information, Communication & Society 1

Klonick K, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018) 131 Harvard Law Review 1598

—, 'The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression' (2020) 129 Yale Law Journal 2418

—, 'Of Systems Thinking and Straw Men Online Speech: Response' (2022) 136 Harvard Law Review Forum 339

Klos M, 'Wrongful Moderation' [2020] Nederlands Juristenblad

Konikoff D, 'Gatekeepers of Toxicity: Reconceptualizing Twitter's Abuse and Hate Speech Policies' [2021] Policy & Internet 502

Koppell JG, 'Pathologies of Accountability: ICANN and the Challenge of "Multiple Accountabilities Disorder"' (2005) 65 Public Administration Review 94

Korchenko O, Korchenko A, Azarov I, 'Analysis of Modern Solutions for the Identification of Anonymous Users', in Dovgyi S, Siemens E, Globa L, Kopiika O,

Stryzhak, O, *Applied Innovations in Information and Communication Technology* (Springer Nature Switzerland 2025).

Korff D, 'The Rule of Law on the Internet and the Wider Digital World' (Council of Europe 2014) Issue Paper

Kosseff J, *The Twenty-Six Words That Created the Internet* (Cornell University Press 2019)

Kosta E and Brewczyńska M, 'Government Access to User Data: Towards More Meaningful Transparency Reports' in Rosa-Maria Ballardini and Olli Pitkänen (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer Law International 2019)

Kosters L and Gstrein OJ, 'TikTok and Transparency Obligations in the EU Digital Services Act (DSA) – A Scoping Review' (2024) 27 Zeitschrift für europarechtliche Studien 110

Kou Y and Gui X, 'Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community', *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2021)

Krasnow EG and Milstead MW, 'FCC Regulation and Other Oxymorons Revisited' (1999) 7 Media Law & Policy 7

Kreimer SF, 'Censorship by Proxy: The First Amendment, Internet Intermediaries, and The Problem of The Weaker Link' (2006) 155 University of Pennsylvania Law Review 11

Krygier M, 'Four Puzzles about the Rule of Law: Why, What, Where? And Who Cares?' (2011) 50 NOMOS: American Society for Political and Legal Philosophy 64

—, 'Tempering Power' in Maurice Adams, Anne Meuwese and Ernst Hirsch Ballin (eds), *Constitutionalism and the Rule of Law* (1st edn, Cambridge University Press 2017)

Kuczerawy A, 'Intermediary Liability & Freedom of Expression: Recent Developments in the EU Notice & Action Initiative' (2015) 31 Computer Law & Security Review 46

—, 'The Power of Positive Thinking Special Issues: Intermediary Liability as a Human Rights Issue' (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 226

—, 'Fighting Online Disinformation: Did the EU Code of Practice Forget about Freedom of Expression?' in Georgios Terzis and others (eds), *Disinformation and Digital Media as a Challenge for Democracy* (1st edn, Intersentia 2019)

—, 'From "Notice and Take Down" to "Notice and Stay Down": Risks and Safeguards for Freedom of Expression' in Giancarlo Frosio (ed), *The Oxford Handbook of Intermediary Liability Online* (Oxford University Press 2019)

—, 'The Good Samaritan That Wasn't: Voluntary Monitoring under the (Draft) Digital Services Act' (*Verfassungsblog*, 12 January 2021) <https://verfassungsblog.de/good-samaritan-dsa/> accessed 21 February 2025

Kuczerawy A and Quintais J, '"Must-Carry", Special Treatment and Freedom of Expression on Online Platforms: A European Story' (2024)

Kukovec D, 'Law and the Periphery' (2015) 21 European Law Journal 406

Land MK, 'Against Privatized Censorship: Proposals for Responsible Delegation' (2020) 60 Virginia Journal of International Law 363

Langvardt K, 'Regulating Online Content Moderation' (2017) 106 The Georgetown Law Journal 1353

—, 'Can the First Amendment Scale?' (2021) 1 Journal of Free Speech Law 273

Latiff R, 'ByteDance's TikTok Cuts Hundreds of Jobs in Shift towards AI Content Moderation' *Reuters* (11 October 2024) <https://www.reuters.com/technology/bytedance-cuts-over-700-jobs-malaysia-shift-towards-ai-moderation-sources-say-2024-10-11/> accessed 27 February 2025

Lee ES and Raymund DR, 'Menu Driven Systems' in Allen Kent and James G Williams, *The Encyclopedia of Microcomputers*, vol 11 (1993)

Leerssen P, 'Cut out by the Middle Man: The Free Speech Implications of Social Network Blocking and Banning in the EU' (2015) 6 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 99

—, 'The Soap Box as a Black Box: Regulating Transparency in Social Media Recommender Systems' (2020) 11 European Journal of Law and Technology 51

—, 'Seeing What Others Are Seeing' (University of Amsterdam 2023)

—, ‘An End to Shadow Banning? Transparency Rights in the Digital Services Act between Content Moderation and Curation’ (2023) 48 Computer Law & Security Review 105790

—, ‘Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act’ (2024) 4 Weizenbaum Journal of the Digital Society

—, ‘Pathways to Private Enforcement of the Digital Services Act (DSA)’ (Institute for Information Law 2025)

Lefouili Y and Madio L, ‘The Economics of Platform Liability’ (2022) 53 European Journal of Law and Economics 319

Lemley M, ‘The Splinternet’ (2021) 70 Duke Law Journal 1397

Liesenfeld A, ‘The Legal Significance of Independent Research Based on Article 40 DSA for the Management of Systemic Risks in the Digital Services Act’ [2024] European Journal of Risk Regulation 1

Light B, Burgess J and Duguay S, ‘The Walkthrough Method: An Approach to the Study of Apps’ (2018) 20 New Media & Society 881

Lobba P, ‘Holocaust Denial before the European Court of Human Rights: Evolution of an Exceptional Regime’ (2015) 26 European Journal of International Law 237

Loughlin M, ‘The Constitutional Imagination: The Constitutional Imagination’ (2015) 78 The Modern Law Review 1

Lyons K, ‘Facebook Reportedly Provided Inaccurate Data to Misinformation Researchers’ (*The Verge*, 11 September 2021) <https://www.theverge.com/2021/9/11/22668396/facebook-provided-inaccurate-data-misinformation-researchers> accessed 16 April 2025

Lyu Y and others, ‘“I Got Flagged for Supposed Bullying, Even Though It Was in Response to Someone Harassing Me About My Disability.”: A Study of Blind TikTokers’ Content Moderation Experiences’, *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2024)

Madiega T, ‘Reform of the EU Liability Regime for Online Intermediaries: Background on the Forthcoming Digital Services Act : In Depth Analysis.’ (European Parliament 2019) PE 649.404

Malcolm J, *Multi-Stakeholder Governance and the Internet Governance Forum* (1st paperback ed, Terminus Press 2008)

Marmor A, 'Two Rights of Free Speech' (2018) 31 Ratio Juris 139

Maroni M, 'The Liability of Internet Intermediaries and the European Court of Human Rights' in Bilyana Petkova and Tuomas Ojanen, *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020)

Marsden C, 'Internet Liability Law Case Study' in Madeleine De Cock Buning and Linda Senden (eds), *Private Regulation and Enforcement in the EU: Finding the Right Balance from a Citizen's Perspective* (Hart Publishing 2020)

Mascheroni G and Holloway D, 'Hybrid Methods for Hybrid Play: A Research Toolkit' in Giovanna Mascheroni and Donell Holloway (eds), *The Internet of Toys: Practices, Affordances and the Political Economy of Children's Smart Play* (Springer International Publishing 2019)

Mast T, 'Platform Law as EU Law' (2024) 73 GRUR International 607

Mauritz FK, 'To Define Is Just to Define' (*Verfassungsblog*, 8 April 2024) <https://verfassungsblog.de/to-define-is-just-to-define/> accessed 20 August 2024

Medzini R and Levi-Faur D, 'Self-Governance via Intermediaries: Credibility in Three Different Modes of Governance' (2023) 25 Journal of Comparative Policy Analysis: Research and Practice 323

Mejtoft T, Hale S and Söderström U, 'Design Friction', *Proceedings of the 31st European Conference on Cognitive Ergonomics* (Association for Computing Machinery 2019)

Meredith S, 'Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal' (*CNBC*, 10 April 2018) <https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html> accessed 18 July 2025

'Meta's New AI System to Help Tackle Harmful Content' (*Meta Newsroom*, 8 December 2021) <https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/> accessed 21 May 2025

Meza RM, Mogoș A-A and Prundaru G, 'Idols of Promotion and Authenticity on TikTok' (2023) 11 Media and Communication

Micova SB and Kukliš Ľ, 'Responsibilities of Video-Sharing Platforms and Their Users' in Heritiana Ranaivoson, Sally Broughton Micova and Tim Raats (eds), *European Audiovisual Policy in Transition* (Routledge 2023)

Montagnani ML, 'A New Liability Regime for Illegal Content in the Digital Single Market Strategy' in Giancarlo Frosio (ed), *The Oxford Handbook of Online Intermediary Liability* (Oxford University Press 2020)

Morrow G and others, 'The Emerging Science of Content Labeling: Contextualizing Social Media Content Moderation' (2022) 73 Journal of the Association for Information Science and Technology 1365

Mueller M, Schmidt A and Kuerbis B, 'Internet Security and Networked Governance in International Relations' (2013) 15 International Studies Review 86

Mustert L, 'Cross-Border Enforcement of the GDPR by Independent Administrative Authorities' (University of Luxembourg 2023)

Myers West S, 'Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms' (2018) 20 New Media & Society 4366

Naslund JA and others, 'Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice' (2020) 5 Journal of Technology in Behavioral Science 245

Nonnecke B and Carlton C, 'EU and US Legislation Seek to Open up Digital Platform Data' (2022) 375 Science 610

Nordemann JB, 'Haftung von Providern im Urheberrecht Der aktuelle Stand nach dem EuGH-Urteil v. 12. 7. 2011 – C-324/09 – L'Oréal/eBay', (2011) 11 GRUR, 977.

Orlando-Salling J, 'The Digital Services Act in the European Periphery: Critical Perspectives on EU Digital Regulation' [2025] European Law Open 1

Orlando-Salling J and Bartolo L, 'The Digital Services Act as Seen from the European Periphery' (*DSA Observatory*, 5 October 2023) <https://dsa-observatory.eu/2023/10/05/the-digital-services-act-as-seen-from-the-european-periphery/> accessed 18 November 2024

Ortolani P, 'If You Build It, They Will Come' [2022] Verfassungsblog <https://verfassungsblog.de/dsa-build-it/> accessed 30 September 2024

Ortutay -Barbara, Ortutay APB and Press A, 'Amnesty Report Finds Facebook Amplified Hate Ahead of Rohingya Massacre in Myanmar' (*PBS News*, 29 September

2022) <https://www.pbs.org/newshour/world/amnesty-report-finds-facebook-amplified-hate-ahead-of-rohingya-massacre-in-myanmar> accessed 18 July 2025

'Outlook on the Global Agenda 2014' (World Economic Forum 2014)

'Overview on Resources Made Available by Member States to Data Protection Authorities and on Enforcement Actions by the Data Protection Authorities' (European Data Protection Board)

Pamuk H, 'Exclusive: Rubio Orders US Diplomats to Launch Lobbying Blitz against Europe's Tech Law' *Reuters* (7 August 2025) <https://www.reuters.com/sustainability/society-equity/rubio-orders-us-diplomats-launch-lobbying-blitz-against-europes-tech-law-2025-08-07/> accessed 12 August 2025

Pascual MG, 'A European Standard of Human Rights Protection?' in Sara Iglesias Sánchez and Maribel González Pascual (eds), *Fundamental Rights in the EU Area of Freedom, Security and Justice* (1st edn, Cambridge University Press 2021)

Pasquale F, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2016)

Peguera M, 'The DMCA Safe Harbors and Their European Counterparts: A Comparative Analysis of Some Common Problems' (2008) 32 Columbia Journal of Law and the Arts 481

Pitruzzella G and Pollicino O, 'The Constitutional Perspective on Freedom of Expression in the Internet Era' in Andrew Spannaus (tr), *Disinformation and hate speech: a European constitutional perspective* (First edition, Bocconi University Press 2020)

Pollicino O, 'Free Speech, Defamation and the Limits to Freedom of Expression in the EU: A Comparative Analysis' in Andrej Savin and Jan Trzaskowski, *Research Handbook on EU Internet Law* (Edward Elgar Publishing 2014)

—, 'European Judicial Dialogue and the Protection of Fundamental Rights in the New Digital Environment: An Attempt at Emancipation and Reconciliation' in Sonia Morano-Foadi and Lucy Vickers (eds), *Fundamental Rights in the EU: A Matter for Two Courts* (Bloomsbury Publishing 2015)

—, 'The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism' [2023] European Law Journal 1

Popper B, 'Airbnb's Worst Problems Are Confirmed by Its Own Data' (*The Verge*, 4 December 2015) <https://www.theverge.com/2015/12/4/9849242/airbnb-data-new-york-affordable-housing-illegal-hotels> accessed 16 April 2025

Pozen DE, 'Straining (Analogies) to Make Sense of the First Amendment in Cyberspace' in David E Pozen (ed), *The Perilous Public Square* (Columbia University Press 2019)

'Preventing "Torrents of Hate" or Stifling Free Expression Online?' (The Future of Free Speech 2024)

'Public Comment by UN Special Rapporteur on Minority Issues Relating to Cases on Hate Speech and Minorities' (*OHCHR*) <https://www.ohchr.org/en/statements-and-speeches/2020/12/public-comment-un-special-rapporteur-minority-issues-relating-cases> accessed 12 April 2025

Purtova N, 'The Law of Everything. Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10 Law, Innovation and Technology 40

Quelle C, 'Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach' (2018) 9 European Journal of Risk Regulation 502

Quintais JP and others, 'Safeguarding User Freedoms in Implementing Article 17 of the Copyright in the Digital Single Market Directive Statements' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 277

Quintais JP, Appelman N and Ó Fathaigh R, 'Using Terms and Conditions to Apply Fundamental Rights to Content Moderation' [2023] German Law Journal 1

Quintais JP, De Gregorio G and Magalhães JC, 'How Platforms Govern Users' Copyright-Protected Content: Exploring the Power of Private Ordering and Its Implications' (2023) 48 Computer Law & Security Review 105792

Quintais JP and Schwemer SF, 'The Interplay between the Digital Services Act and Sector Regulation: How Special Is Copyright?' (2022) 13 European Journal of Risk Regulation 191

Quintel T and Ullrich C, 'Self-Regulation of Fundamental Rights? The EU Code of Conduct on Hate Speech, Related Initiatives and Beyond' in Bilyana Petkova and Tuomas Ojanen, *Fundamental Rights Protection Online* (Edward Elgar Publishing 2020)

Raue B, 'Artikel 22 Vertrauenswürdige Hinweigeber' in Franz Hofmann and others (eds), *Digital Services Act: Gesetz über digitale Dienste* (1. Auflage, Nomos 2023)

Raue B and Heesen H, 'Raue/Heesen: Der Digital Services Act' [2022] Neue Juristische Wochenschrift 3537

Raymond M, 'Puncturing the Myth of the Internet as a Commons' [2013] Georgetown Journal of International Affairs 53

Raymond M and DeNardis L, 'Multistakeholderism: Anatomy of an Inchoate Global Institution' (2015) 7 International Theory 572

Reynolds J, 'The Political Economy of States of Emergency Symposium: Third World Approaches to International Law (TWAIL) Conference: Capitalism and the Common Good' (2012) 14 Oregon Review of International Law 85

Rieder B and Hofmann J, 'Towards Platform Observability' (2020) 9 Internet Policy Review

Riordan R, 'A Case Study of Judicial-Legislative Interactions via the Lens of the DSA's Host Liability Rules' (2025) 2025 10 European Papers – A Journal on Law and Integration 259

Roberts ST, 'Digital Detritus: "Error" and the Logic of Opacity in Social Media Content Moderation' [2018] First Monday <https://www.firstmonday.org/ojs/index.php/fm/article/view/8283> accessed 15 April 2025

'Romanian Authorities Ask EU Commission to Investigate TikTok Role in Election | Euronews' <https://www.euronews.com/next/2024/11/27/romanian-authorities-ask-eu-commission-to-investigate-tiktok-role-in-election> accessed 3 December 2024

'Romanian Authorities Ask EU to Investigate TikTok Role in Election' (*euronews*, 13:47:26 +01:00) <https://www.euronews.com/next/2024/11/27/romanian-authorities-ask-eu-commission-to-investigate-tiktok-role-in-election> accessed 18 December 2024

Romans JT, 'Moral Suasion as an Instrument of Economic Policy' (1966) 56 The American Economic Review 1220

'Roundtable on the Digital Services Act' (Central European Digital Media Observatory 2024)

Rusu C and others (eds), *New Directions in Competition Law Enforcement*, vol 4 (Wolf Legal Publishers 2020)

Salter M and Hanson E, '"I Need You All to Understand How Pervasive This Issue Is": User Efforts to Regulate Child Sexual Offending on Social Media' in Jane Bailey, Asher Flynn and Nicola Henry (eds), *The Emerald International Handbook of Technology-Facilitated Violence and Abuse* (Emerald Publishing Limited 2021)

Sander B, 'Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation' (2020) 43 Fordham International Law Journal 939

Sartor G, 'The Impact of Algorithms for Online Content Filtering or Moderation' (European Parliament 2020)

Savin A, 'The EU Digital Services Act: Towards a More Responsible Internet' (2021) 24 Journal of Internet Law 15

Scanlon TM, 'Why Not Base Free Speech on Autonomy or Democracy Responses' (2011) 97 Virginia Law Review 541

Scheppele KL, 'The Treaties Without a Guardian: The European Commission and the Rule of Law' (2023) 29 Columbia Journal of European Law 93

Schneider J, 'Woher kommt der Wirbel um "Trusted Flagger"?' (*ZDFheute*, 16 October 2024) <https://www.zdf.de/nachrichten/politik/deutschland/trusted-flagger-bundesnetzagentur-zensurvorwurf-100.html> accessed 4 November 2024

Schneider J-P, Siegrist K and Oles S, 'Collaborative Governance of the EU Digital Single Market Established by the Digital Services Act' (2023)

Schütze R, 'From Rome to Lisbon: "Executive Federalism" in the (New) European Union' (2010) 47 Common Market Law Review 1385

Schwemer SF, 'Trusted Notifiers and the Privatization of Online Enforcement' (2019) 35 Computer Law & Security Review 105339

Sekwenz M-T, Wagner B and Parkin S, '"It Is Unfair, And It Would Be Unwise To Expect The User To Know The Law!" – Evaluating Reporting Mechanisms Under The Digital Services Act' <https://data.4tu.nl/datasets/16efb52d-b94d-4180-a5f6-89d93028a8c2> accessed 21 August 2025

—, '"It Is Unfair, and It Would Be Unwise to Expect the User to Know the Law!" – Evaluating Reporting Mechanisms under the Digital Services Act', *Proceedings of the*

2025 ACM Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery 2025)

Senftleben M and Angelopoulos C, 'The Odyssey of the Prohibition on General Monitoring Obligations on the Way to the Digital Services Act: Between Article 15 of the e-Commerce Directive and Article 17 of the Directive on Copyright in the Digital Single Market' [2020] Amsterdam/Cambridge

Senftleben M, Quintais JP and Meiring A, 'How the EU Outsoucres the Task of Human Rights Protection to Platforms and Users: The Case of UGC Monetization' (2023) 38 Berkely Technology Law Journal 101

Shahi, GK, Tessa B, Trujillo A, Cresci S, 'A Year of the DSA Transparency Database: What it (Does Not) Reveal about Platform Moderation During the 2024 European Parliament Election, *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media*, 2025.

Shim Y and Jhaver S, 'Incorporating Procedural Fairness in Flag Submissions on Social Media Platforms' (arXiv, 13 September 2024) <http://arxiv.org/abs/2409.08498> accessed 27 November 2024

Siegrist K, *Systemische Risiken im Digital Services Act* (1. Auflage, Nomos 2025)

Smits JM, 'What Is Legal Doctrine?: On The Aims and Methods of Legal-Dogmatic Research' in Rob Van Gestel, Hans-W Micklitz and Edward L Rubin (eds), *Rethinking Legal Scholarship* (1st edn, Cambridge University Press 2017)

Sørensen E and Torfing J, 'The Democratic Anchorage of Governance Networks' (2005) 28 Scandinavian Political Studies 195

Spano R, 'Intermediary Liability for Online User Comments under the European Convention on Human Rights' (2017) 17 Human Rights Law Review 665

Spindler G, 'The Liability System of Art. 17 DSMD and National Implementation: Contravening Prohibition of General Monitoring Duties' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 344

—, 'Der Referentenentwurf für ein Digitales Dienste Gesetz (DDG-RefE) – good bye TMG und NetzDG: Erster Überblick über die wesentlichen Weichenstellungen' (2023) 39 Computer und Recht 602

Straetmans G and Vereecken J, 'Towards a New Balance Between Private and Public Enforcement in EU Consumer Law' [2024] European Review of Private Law 41

Strowel A and De Meyere J, 'The Digital Services Act: Transparency as an Efficient Tool to Curb the Spread of Disinformation on Online Platforms?' (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 66

Strowel A and Somaini L, 'Towards a Robust Framework for Algorithmic Transparency to Tackle the Dissemination of Illegal and Harmful Content on Online Platforms' (2021) <https://cdn.uclouvain.be/groups/cms-editors-crides/droit-intellectuel/CRIDES_WP_2_2021_Alain%20Strowel%20and%20Laura%20Somaini.pdf> accessed 17 October 2024

Suchman MC, 'Managing Legitimacy: Strategic and Institutional Approaches' (1995) 20 The Academy of Management Review 571

Suzor N, 'Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms' Social Media 11

—, 'What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation' (2019) 13 International Journal of Communication 1526

Szóka DJ Berin, 'The Far Right's War on Content Moderation Comes to Europe | TechPolicy.Press' (*Tech Policy Press*, 11 February 2025) <https://techpolicy.press/the-far-rights-war-on-content-moderation-comes-to-europe> accessed 7 April 2025

Taekema S, 'Theoretical and Normative Frameworks for Legal Research: Putting Theory into Practice' [2018] Law and Method

Take K and others, 'Stoking the Flames: Understanding Escalation in an Online Harassment Community' (2024) 8 Proceedings of the ACM on Human-Computer Interaction 1

'Telegram Still Doesn't Meet Large Platform Requirements under DSA' <https://sg.news.yahoo.com/telegram-still-doesnt-meet-large-095311777.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAGA_0t6lA009T3PlmMHDv1K228kVtDkXBlkuappqM2CuEPAlK-YLwk0wp5XpT96_vaubJGnEV1UPgbwFNEBQsyAQ7o2LyJE8V8Y5na55OlvFvRvSUQEKMWwxHUZzHcvzrMzAVkyJforkxdgsxHWdAI6LoHmo7upC6Icn_8fnfQMi> accessed 1 November 2024

Terzis P, 'Against Digital Constitutionalism' (2024) 3 European Law Open 336

Teubner G, '"And If I by Beelzebub Cast out Devils, …": An Essay on the Diabolics of Network Failure' (2009) 10 German Law Journal 395

—, 'Horizontal Effects of Constitutional Rights in the Internet: A Legal Case on the Digital Constitution' (2017) 3 The Italian Law Journal 193

Teubner G and Golia A, 'Societal Constitutionalism in the Digital World: An Introduction Digital Constitution: On the Transformative Potential of Societal Constitutionalism' (2023) 30 Indiana Journal of Global Legal Studies 1

'The Digital Services Coordinators Database' (*DSC Database*) <https://dscdb.edri.org/> accessed 5 May 2025

Trujillo A, Fagni T and Cresci S, 'The DSA Transparency Database: Auditing Self-Reported Moderation Actions by Social Media' (2025) 9 Proc. ACM Hum.-Comput. Interact. CSCW187:1

Truyens M and Van Eecke P, 'Liability of Domain Name Registries: Don't Shoot the Messenger' (2016) 32 Computer Law & Security Review 327

'Twitch Says Livestream of Buffalo Mass Shooting Was Removed in Less than 2 Minutes | CNN Business' <https://edition.cnn.com/2022/05/15/business/twitch-livestream-buffalo-massacre/> accessed 29 April 2025

Ullrich C, 'Standards for Duty of Care: Debating Intermediary Liability from a Sectoral Perspective' (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 111

Urban JM, Schofield BL and Karaganis J, 'Takedown in Two Worlds: An Empirical Analysis' (2017) 64 Journal of the Copyright Society of the USA 483

Urman A and Makhortykh M, 'How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting across Online Platforms' (2023) 47 Telecommunications Policy 102477

Vaidhyanathan S, *Antisocial Media: How Facebook Disconnects Us and Undermines Democracy* (1st edn, Oxford University Press 2021)

—, 'Facebook and the Folly of Self-Regulation' *Wired* <https://www.wired.com/story/facebook-and-the-folly-of-self-regulation/> accessed 14 April 2025

van Aaken A, 'Constitutional Limits to Nudging: A Proportionality Assessment' in Alexandra Kemmerer and others (eds), *Choice Architecture in Democracies, Exploring the Legitimacy of Nudging* (Hart 2015)

van Cleynenbreugel P, 'Independent Administrative Authorities and the Sectoral "Europeanisation" of National Administrative Law: A European Union Law Perspective' (2023) <https://orbi.uliege.be/bitstream/2268/313933/1/EUDAIMONIA_WP20241.pdf> accessed 17 October 2024

van de Kerkhof J, 'Good Faith in Article 6 Digital Services Act (Good Samaritan Exemption) – The Digital Constitutionalist' (15 February 2023) <https://digi-con.org/good-faith-in-article-6-digital-services-act-good-samaritan-exemption/> accessed 7 April 2025

—, 'Jawboning Content Moderation from a European Perspective' in Charlotte van Oirsouw and others (eds), *European Yearbook of Constitutional Law: Constitutional law in the Digital Era*, vol 5 (1st edn, TMC Asser Press 2024)

—, 'Constitutional Aspects of Trusted Flaggers in The Netherlands' in José van Dijck and others (eds), *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values* (Amsterdam University Press 2025)

—, 'Musk, Techbrocracy, and Free Speech' (*Verfassungsblog*, 17 January 2025) <https://verfassungsblog.de/musk-techbrocracy-and-free-speech/> accessed 24 March 2025

—, 'Article 22 Digital Services Act: Building Trust with Trusted Flaggers' (2025) 14 Internet Policy Review

—, 'The DSA's Tower of Babel: On Digital Services Coordinators and Freedom of Expression' [2025] European Journal of Risk Regulation 1

van de Kerkhof J and Goanta C, 'Shadowbanned on X: The DSA in Action' [2024] European Journal of Risk Regulation 1

van de Kerkhof J and Goanta C, 'Van Code Is Wet Naar Wet Is Code: Geautomatiseerde Transparantie in de Digitaledienstenverordening' [2025] Tijdschrift voor Internetrecht 19

van der Donk B, 'Restricting Access to Legal Content in House Rules: The Platform's Freedom To Conduct a Business as a Counterargument To Limit Users' Freedom of

Expression', *Triangulating Freedom of Speech: Business, Social Rights, and the Freedom of Speech in a Digital Age* (Springer 2021)

van der Donk BBE, 'Should Critique on Governmental Policy Regarding Covid-19 Be Tolerated on Online Platforms? An Analysis of Recent Case-Law in the Netherlands' (2021) 13 Journal of Human Rights Practice 426

van Dijck J, *The Culture of Connectivity: A Critical History of Social Media* (Oxford University Press 2013)

van Drunen MZ, 'The Post-Editorial Control Era: How EU Media Law Matches Platforms' Organisational Control with Cooperative Responsibility' (2020) 12 Journal of Media Law 166

van Hoboken J, 'The Proposed EU Terrorism Content Regulation: Analysis and Recommendations with Respect to Freedom of Expression Implications' (Transatlantic Working Group on Content Moderation Online and Freedom of Expression 2019)

—, 'Hosting Intermediary Services and Illegal Content Online: An Analysis of the Scope of Article 14 ECD in Light of Developments in the Online Service Landscape.' (European Commission 2018)

Vargas Penagos E, 'ChatGPT, Can You Solve the Content Moderation Dilemma?' (2024) 32 International Journal of Law and Information Technology eaae028

Veil W, 'Verpestet ist ein ganzes Land … (Teil I)' (*CR-online.de Blog*, 2024) <https://www.cr-online.de/blog/2024/10/16/verpestest-ist-ein-ganzes-land-teil-i/> accessed 4 November 2024

Wagner B and others, 'Regulating Transparency?: Facebook, Twitter and the German Network Enforcement Act', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) <https://dl.acm.org/doi/10.1145/3351095.3372856> accessed 18 October 2023

'What the DPC-Meta Decision Tells Us about the EU GDPR Dispute Resolution Mechanism' <https://iapp.org/news/a/what-the-dpc-meta-decision-tells-us-about-the-gdprs-dispute-resolution-mechanism> accessed 17 October 2024

Wilman F, Kaleda SK, Loewenthal, P, *The EU Digital Services Act: A Commentary* (Oxford University Press 2013).

Wilman F, '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' (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 317

—, 'The Digital Services Act (DSA): An Overview'

—, 'Between Preservation and Clarification' (*Verfassungsblog*, 2 November 2022) <https://verfassungsblog.de/dsa-preservation-clarification/> accessed 8 May 2025

—, 'De Digital Services Act (DSA): een belangrijke stap naar betere regulering van onlinedienstverlening' (2022) 28 Nederlands tijdschrift voor Europees Recht 220

Wimmers J, 'The Out-of-Court Dispute Settlement Mechanism in the Digital Services Act' (2021) 12 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 381

Wolfsfeld G, Segev E and Sheafer T, 'Social Media and the Arab Spring: Politics Comes First' (2013) 18 The International Journal of Press/Politics 115

Wolters P and Gellert R, 'Towards a Better Notice and Action Mechanism in the DSA' (2023) 14 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 403

Wong D and Floridi L, 'Meta's Oversight Board: A Review and Critical Assessment' (2023) 33 Minds and Machines 261

Woods L, 'Digital Freedom of Expression in the EU' in Sionaidh Douglas-Scott and Nicholas Hatzis, *Research Handbook on EU Law and Human Rights* (Edward Elgar Publishing 2017)

'Working Document Impact Assessment on the Modernisation of the EU Copyright Rules' (European Commission 2016) Commission Staff Working Document SWD(2016) 301 final

Wu FT, 'Collateral Censorship and the Limits of Intermediary Immunity' (2011) 87 Notre Dame Law Review 293

Wu S, 'What Motivates Audiences to Report Fake News?: Uncovering a Framework of Factors That Drive the Community Reporting of Fake News on Social Media' (2024) 12 Digital Journalism 790

Yu PK, 'A Hater's Guide to Geoblocking 2019 Symposium Articles: Intellectual Property in a Globalized Economy: United States Extraterritoriality in International

Business: Copyright Infringement: Geoblocking and the Internet' (2019) 25 Boston University Journal of Science and Technology Law 503

Zhang AQ, Montague K and Jhaver S, 'Cleaning Up the Streets: Understanding Motivations, Mental Models, and Concerns of Users Flagging Social Media Posts' (arXiv, 11 September 2024)

Varia

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on A Digital Single Market Strategy for Europe, COM(2015) 192 final.

European Commission, 'Synopsis report on the public consultation on the regulatory environment for platforms, online intermediaries and het collaborative economy', 2015, accessible via: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15877.

European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Tackling Illegal Content Online, COM(2017)555 final, 28 September 2017.

European Parliament Resolution of 20 October 2020 on the Digital Services Act and fundamental rights issued posed (2020/2022), *OJ C* 404, 6.10.2021.

Code of Conduct on Countering Illegal Hate Speech Online+.

Code of Conduct on Disinformation.

Guidelines on measures to ensure a high level of privacy, safety and security for minors online.

Vele handen maken licht werk: vrijheid van meningsuiting in gezamenlijk platformbestuur in de Digitaledienstenverordening

Contentmoderatie is het proces van het ordenen en aanbevelen van content op het internet en op sociale media. Tussenhandeldiensten, waaronder zeer grote online platforms als Facebook, YouTube, en TikTok, bepalen door contentmoderatie de toon van het online debat en de effectieve grenzen van de uitingsvrijheid. Dit proefschrift onderzoekt hoe de Digitaledienstenverordening ('DSA') een netwerk van actoren toegang biedt tot content moderatie, en hoe dit invloed heeft op de manier waarop de uitingsvrijheid daarbij wordt afgewogen tegen andere rechten en belangen. De DSA stelt eisen aan dit proces en geeft verschillende actoren een formele rol waarmee zij invloed kunnen uit oefenen op het content-moderatie-proces. Die rol wordt onder andere toebedeeld aan gebruikers, het maatschappelijk middenveld, en statelijke actoren. Omdat elk van deze actoren eigen belangen heeft, is het mogelijk dat de uitingsvrijheid in de verdrukking komt. Hoe de DSA het betrekken van deze actoren vorm geeft is daarom van cruciaal belang, en vormt de kern van dit proefschrift.

Hoofdstuk I vormt de inleiding van dit proefschrift. Allereerst wordt daarin de onderzoeksvraag ingeleid: *Hoe beïnvloedt de DSA en de formele rol die zij creëert voor tussenhandeldiensten, gebruikers, het maatschappelijk middenveld, en statelijke actoren het wegen van de uitingsvrijheid in content moderatie?* Deze onderzoeksvraag wordt beantwoord aan de hand van drie relevante groepen van actoren: tussenhandeldiensten, gebruikers en het maatschappelijk middenveld, en statelijke actoren. Bij deze beantwoording is gebruik gemaakt van verschillende methoden. Naast zuiver juridisch onderzoek wordt gebruik gemaakt van empirische onderzoeksmethoden uit de computer- en mediawetenschappen. Dit is nodig om te beschrijven hoe content moderatie werkt, en hoe tussenhandeldiensten zich conformeren aan de DSA.

Naast de methodologie schetst hoofdstuk I de achtergrond van het proefschrift. De DSA is een relatief nieuwe Europese verordening die een geharmoniseerd regime creëert voor tussenhandeldiensten, waaronder online platforms en online zoekmachines. De DSA volgt de Richtlijn elektronische handel uit 2000 op. Daarmee vormt zij een veel gezochte wettelijke interventie in het online domein: het internet is sinds 2000 fors veranderd, en brengt nieuwe gevaren met zich mee. Er bestaat een

lacune tussen die gevaren en de instrumenten die bestaande wettelijke instrumenten bieden om die gevaren het hoofd te bieden. Hoofddoelen van de DSA zijn het voorkomen van fragmentatie tussen lidstaten, het stimuleren van verantwoordelijk en zorgvuldig gedrag van tussenhandeldiensten om een veilige, voorspelbare, en betrouwbare online omgeving van Unieburgers te garanderen. Deze doelen beoogt de DSA te realiseren met een reeks zorgvuldigheidsverplichtingen die transparantie vergroten en die het mogelijk maken dat tussenhandeldiensten rekenschap[1] afleggen over hun handelen. Deze zorgvuldigheidsverplichtingen volgen een gelaagde structuur: naarmate de tussenhandeldienst groter wordt – en daarmee haar maatschappelijke impact – zijn meer verplichtingen van toepassing. In deze constellatie zijn de meeste verplichtingen van toepassing op Zeer Grote Online Platforms en Zeer Grote Online Zoekmachines.

Een deel van de zorgvuldigheidsverplichtingen rondom content moderatie bestaat uit het creëren van een netwerk van actoren. Dit kan worden gezien in het licht van Robyn Caplan's theorie over *networked platform governance*.[2] Zij omschrijft hoe platformbestuur gedelegeerd kan worden naar meerdere actoren, die in een netwerk gezamenlijk het bestuur van een platform bepalen. Deze actoren hebben allemaal verschillende machtsposities en verschillende mate van invloed. Dit maakt het moeilijk om te bepalen aan wie het resultaat van dat platformbestuur – hoe content gemodereerd wordt, en daarmee hoe de uitingsvrijheid gewogen wordt – uiteindelijk kan worden toegerekend. Deze spanning wordt onderzocht in de volgende hoofdstukken.

Hoofdstuk II geeft een empirische achtergrond inzicht in hoe het content moderatie landschap eruit ziet. Het hoofdstuk onderzoekt hoe de DSA transparantie database inzicht geeft in content moderatie praktijken, en of tussenhandeldiensten inderdaad transparantie verschaffen. Bevindingen uit een empirische analyse van de database laten zien dat, hoewel de architectuur van de database geschikt lijkt om transparantie te verschaffen, de invulling door tussenhandeldiensten soms te wensen over laat. Opvallend is dat het gros van de content moderatiebeslissingen buiten sociale mediaplatforms plaatsvindt. Vrijwel alle content moderatiebeslissingen worden gemaakt op basis van algemene voorwaarden, niet op basis van geldend recht. Daarnaast valt op dat tussenhandeldiensten verschillende moderatiewijzen hanteren. Sommigen rapporteren dat zij alles

[1] Rekenschap wordt hier gebruikt als vertaling van het woord 'accountability'.

[2] Losjes vertaald zou dit 'platformbestuurnetwerk' of gezamenlijk platformbestuur kunnen betekenen.

geautomatiseerd doen, terwijl anderen stellen dat alle moderatie door menselijke moderatoren gebeurt. Op de waarde van de verschafte informatie valt veel af te dingen: het is in veel gevallen onduidelijk waarom content verwijderd is, omdat tussenhandeldiensten gebruikmaken van generieke verwijzingen naar algemene voorwaarden. De belangrijkste conclusies van dit hoofdstuk voor het proefschrift zijn dat tussenhandeldiensten grotendeels modereren op basis van hun eigen algemene voorwaarden, en dat de mate van kennisdeling tussen actoren grotendeels bepaald wordt door de mate waarin tussenhandeldiensten kennis verschaffen.

Hoofdstuk III richt zich op de aansprakelijkheidsuitzondering die tussenhandeldiensten genieten onder de DSA. De academische consensus op dit onderwerp is dat deze niet veranderd is sinds de Richtlijn elektronische handel en dat de zorgvuldigheidsverplichtingen in de DSA los gezien moeten worden van de aansprakelijkheidsuitzondering. Hoofdstuk III betoogt dat die conclusie genuanceerd kan worden. Ten opzichte van de Richtlijn elektronische handel kan in sommige gevallen wel degelijk een hogere mate van zorgvuldigheid verwacht worden, waardoor tussenhandeldiensten eventueel kunnen worden uitgesloten van de aansprakelijkheidsuitzondering.

Een strengere interpretatie van de aansprakelijkheidsuitzondering kan in sommige gevallen slachtoffers van illegale content beter in staat stellen de tussenhandeldienst aansprakelijk te houden voor het hosten van illegale content. De strengere interpretatie kan als aanmoediging dienen om een beter *notice-and-action* mechanisme te faciliteren, moderatiepraktijken aan te scherpen tegen illegale content, en een beter evenwicht te bewerkstelligen tussen de belangen van de tussenhandeldienst en die van de benadeelde persoon. De uitingsvrijheid zoals deze geïnterpreteerd wordt door het Europees Hof voor de Rechten van de Mens (EHRM) hoeft zich niet tegen die interpretatie te verzetten. Evenmin zou deze interpretatie niet gedragen worden door de DSA. Artikel 16 voorziet dat hosting diensten een eenvoudig toegankelijk en gebruiksvriendelijk *notice-and-action* mechanisme aan moeten bieden. Het niet aanbieden van zo'n mechanisme verstoort de mogelijkheden om hosting diensten aansprakelijk te houden. Om te voorkomen dat het loont om een disfunctioneel mechanisme aan te bieden, zou een hosting dienst die zich aan de verplichtingen van artikel 16 onttrekt uitgesloten kunnen worden van de aansprakelijkheidsuitzondering. Hetzelfde geldt voor online platforms die zich structureel onttrekken aan de verplichtingen van artikel 23 DSA. Artikel 23 verplicht online platforms tot het blokkeren van gebruikers die frequent kennelijk illegale content uploaden. Wanneer een online platform dat niet doet, handelt het

online platform niet te goeder trouw, en kan het daarmee uitgesloten worden van de aansprakelijkheidsuitzondering van artikel 6 DSA.

Het hoofdstuk betoogt ook dat met aangescherpte zorgvuldigheidsverplichtingen ten aanzien van kennelijk illegale inhoud, terroristische inhoud, en kindermisbruik, de lat verlaagd kan worden ten opzichte van de Richtlijn elektronische handel voor feitelijke toetsing of een hosting dienst op de hoogte is van illegale content en daarmee onder verplichting staat die te verwijderen. Dit zou ook kunnen leiden tot het uitsluiten van de aansprakelijkheidsuitzondering van te kwader trouw handelende actoren. Het hoofdstuk betoogt dat tussenhandeldiensten zo een grotere mate van rekenschap moeten afleggen over hoe zij hun content moderatieprocessen inrichten.

Hoofdstuk IV richt zich op een specifiek aspect uit het aansprakelijkheidsvraagstuk, namelijk het *notice-and-action* mechanisme. Het *notice-and-action* mechanisme vervult een sleutelrol in het aansprakelijk houden van hosting diensten voor illegale content, en maakt het voor gebruikers mogelijk om hun voorkeuren kenbaar te maken en invloed uit te oefenen op content moderatie. Het hoofdstuk onderzoekt met een empirische *walkthrough* hoe AliExpress, TikTok, Twitch, en X hun *notice-and-action* mechanismen vormgeven, of dit strookt met de vereisten van artikel 16 DSA, en of dit eventuele aansprakelijkheid van hostingdiensten beïnvloedt. Het onderzoek laat zien dat de genoemde tussenhandeldiensten hun *notice-and-action* mechanismen verschillend vormgeven, en dat die vormgeving vaak geijkt is op de dienst die zij aanbieden. Ook laat het hoofdstuk zien dat met name bij de gemakkelijke toegankelijkheid grote verschillen zichtbaar zijn tussen de tussenhandeldiensten. Dit komt doordat de *notice-and-action* mechanismen voor de DSA vaak sterk verweven zijn met de *notice-and-action* mechanismen die worden gebruikt bij schendingen van de algemene voorwaarden, ook al maakt – en vereist – de DSA daartussen een scherp onderscheid. Deze sterke verwevenheid kan er ook toe leiden dat gebruikers content rapporteren buiten het mechanisme van artikel 16 om.

Het hoofdstuk concludeert dat de vormgeving van de bestaande mechanismen misleidend kan zijn voor gebruikers. Bovendien beïnvloedt de vormgeving de aansprakelijkheid voor eventuele illegale content, zeker nu mechanismen voor het rapporteren van inhoud onder algemene voorwaarden vaak niet de gelegenheid bieden inzicht te verschaffen in waarom bepaalde content illegaal wordt geacht. Wanneer gebruikers echter niet voldoende toelichting verschaffen over de kwalificatie van bepaalde content als 'illegaal', hoeven hosting diensten de content niet te verwijderen.

Het hoofdstuk laat verder zien dat naleving van artikel 16 wisselend is, en dat er bij een aantal tussenhandeldiensten ruimte voor verbetering is. Omdat *notice-and-action* mechanismen voor gebruikers een belangrijke weg zijn om hun voorkeuren kenbaar te maken – en daarmee bij te dragen aan content moderatie en het afwegen van de uitingsvrijheid – is het nodig dat artikel 16 strenger wordt gehandhaafd.

Hoofdstuk V verkent artikel 22 DSA, dat gaat over betrouwbare flaggers (*'trusted flaggers'*). Betrouwbare flaggers kunnen statelijke of niet-statelijke entiteiten zijn, zoals NGOs, ministeries, toezichthouders of de politie. Betrouwbare flaggers hebben een geprivilegieerde positie in het content moderatie proces; hun kennisgevingen dienen met prioriteit en onverwijld te worden behandeld. Betrouwbare flaggers opereren primair in een bilaterale verhouding tot het online platform, wat leidt tot ondoorzichtigheid van hun handelen. Daarnaast is het waarschijnlijk dat online platforms druk voelen om mee te gaan in verwijderverzoeken van betrouwbare flaggers door de gepriviligeerde positie die zij hebben. Dit kan leiden tot het verwijderen van legale content, hetgeen de uitingsvrijheid onder druk zet. Door de ondoorzichtigheid van hun opereren kan dit slecht gecorrigeerd worden.

Artikel 22 DSA formaliseert dit proces. Betrouwbare flaggers worden benoemd door de nationale digitaledienstencoördinator. Zij krijgen met hun status toegang tot álle online platforms – dus niet alleen die waarmee ze een bilaterale verhouding hebben. Artikel 22 maakt het raamwerk waarin betrouwbare flaggers opereren bovendien doorzichtiger. Zij mogen enkel kennisgeven van illegale inhoud aan online platforms, wat betekent dat ze binnen de grenzen van de nationale- en uniewetgeving moeten opereren. Daarnaast worden ze geacht te rapporteren over al hun kennisgevingen, hetgeen overmatig gebruik van hun privileges tegengaat. Daarmee wordt een deel van de kritiek op het oorspronkelijke raamwerk van betrouwbare flaggers beantwoord.

De regulering via artikel 22 DSA roept ook vragen op. Betrouwbare flaggers worden benoemd door nationale autoriteiten, en hanteren vaak nationaal recht. De bepaling bevat weinig waarborgen om te zorgen dat nationaal recht niet Uniebreed wordt uitgerold, hetgeen ook de uitingsvrijheid onder druk zou kunnen zetten. Om deze risico's te bestrijden bepleit het hoofdstuk een geharmoniseerd portaal dat betrouwbare flaggers gebruiken om inhoud te rapporteren.

Een ander zwak punt van artikel 22 DSA is dat het niet duidelijk is wat het onafhankelijkheidsvereiste betekent in de benoeming van betrouwbare flaggers. Hoewel dit bedoeld is om geldstromen vanuit tussenhandeldiensten tegen te gaan, is

het in de praktijk vaak zo dat betrouwbare flaggers daar financieel van afhankelijk zijn. Het hoofdstuk doet op dit punt de suggestie dat de Europese Commissie kan faciliteren dat betrouwbare flaggers een faire financiële tegemoetkoming ontvangen ter ondersteuning van hun taak in het tegengaan van illegale inhoud.

Hoofdstuk VI onderzoekt hoe handhaving door digitaledienstencoördinatoren zich verhoudt tot de uitingsvrijheid. Digitaledienstencoördinatoren zijn nationale toezichthouders op naleving van de DSA. In hun toezichthoudende rol hebben zij mogelijkerwijs veel impact op de uitingsvrijheid op het internet; zij mogen content moderatiepraktijken sanctioneren, tussenhandeldiensten beboeten, en zelfs toegang in het algemeen blokkeren. Hun rolperceptie heeft daarmee een belangrijke invloed op de toegankelijkheid van bepaalde content, en dus op de uitingsvrijheid.

De handhavingsstructuur voor digitaledienstencoördinatoren is complex. Digitaledienstencoördinatoren zijn in beginsel onafhankelijke instituties, maar worden vaak ondergebracht bij reeds bestaande toezichthouders, zoals media- en consumentenautoriteiten. Dit fragmenteert het toezichtlandschap in Europa. Dat is problematisch, omdat de bevoegdheid van de digitaledienstencoördinator niet verder rijkt dan de tussenhandeldiensten die binnen de landsgrenzen gevestigd zijn. Een complicerende factor is bovendien dat digitaledienstencoördinatoren hun handhavingstaak delen met de Europese Commissie, die bevoegd is toe te zien en te handhaven ten aanzien van Zeer Grote Online Platforms en Zeer Grote Online Zoekmachines. Dit betekent dat samenwerking vereist is.

Samenwerking in grensoverschrijdende kwesties en in Unieverband werpt echter problemen op. Het eerste probleem is dat illegaliteit van inhoud nationaal én unierechtelijk bepaald is. Hoewel enige overeenstemming bestaat binnen de Europese Unie over wat illegaal is, lopen de opvattingen van lidstaten op sommige vlakken sterk uiteen. De Hongaarse anti-LHBTIQ+-wetten zijn hierbij een pregnant voorbeeld. Hoewel illegaliteit van bepaalde uitingen mede moet worden getoetst aan de hand van het Handvest, is het onduidelijk of toezichthouders in staat zijn deze beoordeling goed te maken, nog los van of de vraag of tussenhandeldiensten dit zelf goed doen. Dit zet de uitingsvrijheid mogelijk onder druk.

Een tweede probleem is dat de bevoegdheidsstructuur van de DSA maakt dat handhaving zich sterk concentreert in een klein aantal lidstaten, met name Ierland en Nederland. Veel klachten moeten door de Ierse of de Nederlandse toezichthouder behandeld worden, wat operationele knelpunten kan veroorzaken. De

uitingsvrijheid op het internet kan onder druk komen te staan wanneer nationale toezichthouders vastlopen.

Als laatste complicatie in de handhavingsstructuur van de DSA wijst het hoofdstuk erop dat nationale handhavingsstrategieën sterk uiteen kunnen lopen, juist als gevolg van de huisvesting van de digitaledienstencoördinatoren bij toezichthouders uit verschillende rechtsgebieden. Dit geldt eens te meer wanneer middelen tekortschieten bij toezichthouders in kleinere lidstaten. Ook hierdoor kan de bescherming van de uitingsvrijheid onder druk komen te staan.

Het hoofdstuk concludeert dat nagedacht moet worden over mechanismen waardoor toezichthouders tot actie gedwongen kunnen worden, over horizontale prejudiciële procedures tussen lidstaten om interpretaties van nationaal recht te toetsen, en over herverdeling van de toezichtsvergoeding van artikel 43 DSA om kleine lidstaten te ondersteunen. Op deze wijze kan de uitingsvrijheid minder onder druk komen te staan in handhaving van de DSA door digitaledienstencoördinatoren.

Hoofdstuk VII richt zich op inmenging van statelijke actoren in content moderatie. Wanneer een statelijke actor zonder formele bevoegdheid druk uitoefent op tussenhandeldiensten om over te gaan tot het modereren van bepaalde types content, kan dit worden gezien als *jawboning*. *Jawboning* is een term die gebruikt wordt in de (voornamelijk Amerikaanse) literatuur om te beschrijven hoe een actor buiten zijn bevoegdheid om een beleidsdoel probeert te bereiken door middel van het uitoefenen van (informele) druk op private actoren. Wanneer dit wordt gedaan om content moderatie te beïnvloeden knelt dit met de uitingsvrijheid. Immers, beperkingen van de uitingsvrijheid moeten gerechtvaardigd worden, wat betekent dat beperkingen een wettelijke basis moeten hebben, een legitiem doel moeten nastreven, en proportioneel en nodig zijn om dit doel te realiseren. Dit kan als geheel niet getoetst worden wanneer de beperking op informele wijze plaatsvindt.

Een actueel voorbeeld van *jawboning* in content moderatie zijn richtlijnen als de gedragscode tegen *hate speech*. In beginsel worden dergelijke richtlijnen en gedragscodes voorgesteld als vormen van zelf- of co-regulering. Vaak is het echter zo dat een wetgever private actoren sterk aanmoedigt om hieraan deel te nemen, onderhandelingen faciliteert, of zelfs dreigt met verregaande wetgeving om de beleidsdoelen die ten grondslag liggen aan de gedragscodes te bereiken. Dit is ook te zien bij de gedragscode tegen *hate speech*. De Commissie gaf expliciet te kennen dat wanneer de gedragscode niet voldoende zou worden nageleefd, strengere

wetgeving verwacht kon worden. Zij stuurde daarbij aan op het meer proactief tegengaan van *hate speech*.

Een dergelijke aanpak schuurt met de uitingsvrijheid, in het bijzonder omdat de grenzen van *hate speech* vaag en bijzonder context-gevoelig zijn. Aanmoediging tot meer proactieve moderatie kan leiden tot onnodige inperking van de uitingsvrijheid van gebruikers. Dit zou dan het gevolg zijn van het handelen van de Commissie, wat aan het EVRM en het Handvest getoetst zou moeten worden. Dit is echter lastig omdat het vaak gaat om informele druk. Het hoofdstuk roept daarom op om transparant te zijn over wat de inbreng is van statelijke actoren in co- of zelf-regelgevende processen, opdat hun handelen beter getoetst kan worden.

Hoofdstuk VIII brengt de bevindingen van de verschillende hoofdstukken samen. Het schetst hoe de interacties die de DSA creëert tussen de verschillende betrokken actoren spanningen kunnen oproepen, en dat niet duidelijk is hoe in contentmoderatieprocessen de uitingsvrijheid wordt gewogen. In termen van Caplan's theorie over gedeeld platformbestuur betekent dit dat weliswaar alle actoren invloed hebben op content moderatie, maar dat onduidelijk blijft hoe deze invloed zich verhoudt tot het wegen van de uitingsvrijheid. Het hoofdstuk roept aan de hand van de suggesties die in de afzonderlijke artikelen zijn gedaan op tot een versterking van de mogelijkheden en verplichtingen van de individuele actoren om rekenschap af te leggen over hun rol in contentmoderatieprocessen. Daarbij moet ook worden verduidelijkt hoe hun inbreng de uitingsvrijheid beïnvloedt. Naast deze verbeteringen is het nodig om duidelijkheid te creëren over wat de DSA wél en wat de DSA niét kan bewerkstelligen, en er daarmee voor te zorgen dat het geen Zwitsers zakmes zonder juridische basis wordt. Alleen op die manier kunnen een bestendige uitvoering en handhaving van de DSA gegarandeerd worden.

Curriculum Vitae

Jacobus Johannes Wilhelmus van de Kerkhof (12 October 1995) was born in Millingen aan de Rijn, the Netherlands. He obtained his Bachelor of Arts in liberal arts and sciences *cum laude* at University College Roosevelt in 2015, having taken courses in antiquities, linguistics, law, history, and rhetorics. In 2017 he obtained his LLB from Utrecht University, having followed the 'Utrecht Law College' honours program.

Between 2015 and 2019 Jacob gained industry experience in intellectual property law and corporate law during subsequent employment at Ventoux Advocaten, Van Benthem & Keulen, and De Brauw Blackstone Westbroek.

In 2020, he completed his LLM degree in Legal Research from Utrecht University having written a thesis on 'Standard Essential Patent injunctions post *Huawei v ZTE*: a comparative review of the Netherlands, Germany, and the United Kingdom' under prof. dr. mr. Peter Blok. In 2021, he graduated *cum laude* from the LLM International Trade and Investment Law at the University of Amsterdam having written a thesis on 'Environmental Counterclaims after *Burlington & Perenco*: the Mirage of Progress' under dr. Andrea Leiter.

In 2021, Jacob started his PhD at the Montaigne Centre for Rule of Law & Administration of Justice at Utrecht University. His research is supervised by prof. dr. mr. Peter Blok, prof. dr. mr. Janneke Gerards, and supervised daily by dr. Catalina Goanta. His research focuses on freedom of expression in the context of the Digital Services Act. As of 1 November 2025, Jacob is employed as associate (advocaat) at De Brauw Blackstone Westbroek.

From 2019 until 2025, Jacob was part of the Dutch Olympic rowing team, winning a silver medal at the Paris 2024 Olympic Games in the men's eight.

PhD Portfolio

PhD Articles

1. Van de Kerkhof, J., 'The Good, the Bad, the Compliant: Notice-and-Action Mechanisms under Article 16 DSA'.

2. Van de Kerkhof, J., 'To Host or not to Host: Rethinking Intermediary Liability Exemption in the Digital Services Act', *Nordic Journal of European law,* 2025.

3. Van de Kerkhof, J., 'The DSA's Tower of Babel: on the Digital Services Act and Freedom of Expression', *European Journal of Risk Regulation,* 2025.

4. Van de Kerkhof, J., 'Article 22 DSA: Building Trust with Trusted Flaggers', *Internet Policy Review,* 2025.

5. Van de Kerkhof, J., 'Jawboning Content Moderation from a European Perspective', in: van Oirsouw, C., de Poorter, J., Leijten, I., van der Schyff, G., Stremler, M., & de Visser, M. (eds.), *European Yearbook of Constitutional Law 2023: Constitutional Law in the Digital Era,* the Hague: T.M.C. Asser Press, 2024.

6. Kaushal, R., van de Kerkhof, J., Goanta, C., Spanakis, G., & Iamnitchi, A., 'Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database', *Proceedings of the 2024 ACM Conference on Fairness, Accountability and Transparency,* 2024.

Other publications

1. Goanta, C., Zannettou, S., Kaushal, R., van de Kerkhof, J., Bertaglia, T., Annabell, T., Gui, H., Spanakis, G., & Iamnitchi, A., 'the Great Data Standoff: Researchers vs Platforms under the Digital Services Act', *Proceedings of the 20th International AAAI Conference on Web and Social Media,* 2026.

2. Van de Kerkhof, J., 'Article 22', in Twigg-Flesner, C., & Woods., L. (eds.), *Digital Services Act Commentary,* Cheltenham: Edward Elgar Publishing, forthcoming.

3. Van de Kerkhof, J., 'Musk, Techbrocracy and Free Speech', in: Alemanno, A. & Veraldi, J. (eds.), *Musk, Power, and the EU: Can EU Law Tackle the Challenges of Unchecked Plutocracy,* Berlin: Verfassungsbooks, 2025.

4. Annabell, T., Kelder, T., van de Kerkhof, J., Gui, H., & Goanta, C., 'Influencer

as individual and trader: exploring the boundaries of discrimination in influencer marketing from a multidisciplinary perspective', *Continuum: Journal of Media & Cultural Studies,* 2025.

5. Annabell, T., Gorwa, R., Scharlach, R., van de Kerkhof, J., Bertaglia, T., 'TikTok Search Recommendations: Governance and Research Challenges', *Workshop Proceedings of the 19th International AAAI Conference on Web and Social Media,* 2025.

6. Van de Kerkhof, J., 'Constitutional Aspects of Trusted Flaggers in the Netherlands', in: van Dijck, J., van Es, K., Helmond, A., van der Vlist, F. (eds.), *Governing the Digital Society: Platforms, Artificial Intelligence, and Public Values,* 2025.

7. Van de Kerkhof, J. & Goanta, C., 'Van Code is Wet naar Wet is Code: Geautomatiseerde Transparantie in de Digitaledienstenverordening', *Tijdschrift voor Internetrecht,* 2025.

8. Van de Kerkhof, J., & Goanta, C., 'Shadowbanned on X: The DSA in Action', *European Journal of Risk Regulation,* 2025.

9. Van de Kerkhof, J., & Goanta, C., 'Digital Constitutionalism and Consumer Protection', in De Gregorio, G., Pollicino, O., & Valcke, P. (eds.), *The Oxford Handbook of Digital Constitutionalism,* Oxford: Oxford University Press, 2024.

10. Goanta, C., Straub, M., van de Kerkhof, J. (eds.) *Radical Reforms: Bringing Fairness to Social Media Contracts,* Berlin: Verfassungsbooks, 2023.

Web-based output

1. Van de Kerkhof, J., 'Unpacking the EU's Digital Services Act Delegated Regulation on Data Access', *TechPolicyPress*, 2025.

2. Van de Kerkhof J., 'Alexandru Patrascu t. Roemenië en Klimova e.a. t Rusland: Verheldering over uitingsvrijheid en aansprakelijkheid online content van derden', *EHRC Updates,* 2025

3. Van de Kerkhof, J., 'Peradze v Georgia: vulgar language, public morals, and the right to peaceful assembly, *Strasbourg Observers,* 2023.

4. Van de Kerkhof, J., 'Sanchez v France: the expansion of intermediary liability in the context of online hate speech, *Strasbourg Observers,* 2023.

5. Van de Kerkhof, J., 'Good Faith in Article 7 Digital Services Act (Good Samaritan Exception)', *The Digital Constitutionalist,* 2023.

6. Van de Kerkhof, J., 'Being Misled by a Google Ad: Who is Responsible?', *The Digital Constitutionalist,* 2023.

7. Van de Kerkhof, J., & Goanta, C., 'Fakeness in Political Popularity', *Verfassungsblog,* 2022.

8. Van de Kerkhof J., 'Dutch District Courts on the Shopping Mall Analogy in content moderation and the Proposed Digital Services Act', *The Digital Constitutionalist,* 2022.

9. Van de Kerkhof, J., 'Welke rol spelen Big Tech-bedrijven als tussenpersoon voor adverteerders', *Montaigne Centre Blog,* 2022.

10. Van de Kerkhof, J., 'Biancardi v Italy, a Broader Right to be Forgotten', *Strasbourg Observers*, 2022.

11. Van de Kerkhof, J.,'Elon Musk wants to Buy Twitter to Create a Free Speech Utopia: Now What?', *Verfassungsblog,* 2022.

Policy Input

1. Aade, L., Annabell, T., Bertaglia, T., Goanta, C., Istrate, S., van de Kerkhof., J., Sax, M., 'Digital Fairness Act Public Consultation', European Commission, 2025.

2. Aguerre, C., Ahmed, A., Archegas, J.V., Arias-Ortiz, W.A., Botero Arcila, B., Bottino, C., Brossi, L., Cala Vitery, E., Castillo, A.M. Cortesi S., Fieseler C., Garcia Avis, I., Gasser, U., Gomez Ayerbe, P., Guio Espanol, A., Hernandez-Serret, J., Hidalgo, J., Lemos, R., Pilar Llorens, M., Londono, F.C., Lutz, C., Marco, F., Mejia, M.I., Molnar, A., Munoz, V., Roca Trenchs, N., Schönborn, S., Sigaud-Sellos, P., Siewert, M., Steibel, F., Affonso Souza, C., van de Kerkhof, J., 'Inverting the Brussels Effect: What the EU can learn from Latin America in Digital Governance', TUM Think Thank, 2025.

3. Annabell, T., Bertaglia, T., Goanta, C., & van de Kerkhof, J., 'Public Consultation on the Draft Committee of Ministers Recommendation on Online Safety and Empowerment of Content Creators and Users', Council of Europe, 2025.

4. Van de Kerkhof, J., De Gregorio, G., De Meyere, J., Gentile, G., & Kuczerawy, A., 'Policy Brief for the European Commission's Digital Services Act Delegated Act on Data Access Consultation', European Commission, 2024.

5. Van de Kerkhof, J., Annabell, T., Goanta, C., Iamnitchi, A., Kaushal, R., & Spanakis, G., 'Policy Brief for the European Commission's Digital Services Act Delegated Act on Data Access Consultation', European Commission, 2024.

6. Aade, L., Annabell, T., Bertaglia, T., Iamnitchi, A., van de Kerkhof, J., Goanta, C., & Gui, H., 'Policy Brief for the European Commission's Digital Services Act implementation consultation', European Commission, 2023.

7. Aade, L., van de Kerkhof, J., Goanta, C., 'Policy Brief for French Government Public Consultation on Influencer Regulation', France, 2023.

Invited Participation

1. 'ERC workshop on data access for research on systemic risks under Digital Services Act (DSA) Art 40', 22-23 October 2025, European Commission, Belgium.

2. 'The DSA's Tower of Babel: on the Digital Services Coordinator and Freedom of Expression', 14 October 2025, AGCOM: the EU Regulation on Vloggers and Influencers, hybrid.

3. 'The Good, the Bad, the Compliant', 9 October 2025, Academic Briefing, European Commission Representation in the Netherlands, the Netherlands.

4. 'It Takes a Village to Moderate a Platform', 7 October 2025, European University Institute, Italy.

5. 'Constitutional Aspects of Trusted Flaggers in the Netherlands', 22 September 2025, Governing the Digital Society Book Launch , Utrecht University, the Netherlands.

6. 'The Digital Services Act and Consumer Protection', 29 August 2025, Forbrukertilsynet, Norway.

7. 'It Takes a Village to Moderate a Platform', 28 August 2025, BI Norwegian Business School, Norway.

8. 'Hemispheres Forum', 7-9 July 2025, Hemispheres, Technische Universität

München, Germany.

9. ‘The DSA’s Tower of Babel: on the Digital Services Coordinator and Freedom of Expression’, 3 & 4 July 2025, Young European Law Scholars Conference, Utrecht University, the Netherlands.

10. ‘The Great Data Standoff: Researchers vs Platforms under the Digital Services Act’, 30 June, European Workshop on Algorithmic Fairness, TU Eindhoven, the Netherlands.

11. ‘The DSA’s Tower of Babel: on the Digital Services Coordinator and Freedom of Expression’, 16 & 17 June 2025, International Conference of Jean Monnet Chair dig_INFlow, Adam Mickiewicz University, Poland.

12. ‘Automated Transparency: A Legal and Empirical Analysis of the Digital Services Act Transparency Database’, 3 & 4 June 2025, Association of Internet Researchers Flashpoint Symposium, Universität Bremen, Germany.

13. ‘The DSA’s Tower of Babylon: on the Digital Services Coordinator and Freedom of Expression’, 22 & 23 May 2025, ICON-S Benelux, Université Libre Bruxelles, Belgium.

14. ‘Techbrocracy and Free Speech’, 10 April 2025, Disinformation, Social Media, and the Democratic Publish Sphere, Université de Brest, France.

15. ‘Risky Business, A Critical Analysis of Systemic Risks in the DSA’, 16-17 January 2025, Governing or governed by online platforms? Interdisciplinary perspectives on Constitutionalism, the EU, and Corporate Power, Maastricht University, the Netherlands.

16. ‘Automated Transparency: a Legal and Empirical Analysis of the Digital Services Act Transparency Database’, 10 December 2024, Digital Speech and Rights: A Workshop on Automated Content Moderation, Örebro University, Sweden.

17. ‘Automated Transparency: a Legal and Empirical Analysis of the Digital Services Act Transparency Database’, 25 October 2024, DSA Transparency Database Workshop, European Commission, hybrid.

18. ‘Enforcing Fundamental Rights Through National Administrative Authorities: The Case of the Digital Services Coordinator’, 4 October 2024,

RESHUFFLE Young Scholars Workshop, KU Leuven, Belgium.

19. 'Art. 22 DSA: Building Trust with Trusted Flaggers', 8 March 2024, HUMANads Outreach Event, Utrecht University, the Netherlands.

20. 'Art. 22 DSA: Building Trust with Trusted Flaggers', 15 February 2024, The DSA and Platform Regulation Conference, University of Amsterdam, the Netherlands.

21. 'Monetising Political Speech – Enforcement under the Digital Services Act', 26 October 2023, ICON-S Benelux, Maastricht University, the Netherlands.

22. 'Good Faith in Article 7 Digital Services Act (Good Samaritan Exemption)', 21 November 2022, DigiCON, European University Institute, Italy.

23. 'Jawboning in European Content Moderation', 17 November 2022, Masterclass: prof. Lilian Edwards on regulation of technology, Utrecht University, the Netherlands.

